Development of a Pathogen Profiling Approach for Detecting and Dissecting Markers of Pathogenicity and Hyper-Variability in Group B Streptococci

Thesis

Development of a pathogen profiling approach for detecting and dissecting markers of pathogenicity and hyper-variability in group B streptococci.

R.P. Loy BSc.

Submitted for Doctorate of Philosophy

Faculty of Life Sciences

The Open University

September 2012

DATE of SUBMISSION : 25 SEPTEMBER 2012
DATE of AWARD : 24 JANUARY 2013

ProQuest LLC.
789 East Eisenhower Parkway
P.O. Box 1346
Ann Arbor, MI 48106 – 1346

Acknowledgments

This is probably the most overused cliché in the writing of acknowledgments but this section of the thesis is both the easiest and hardest section to write with so many people to thank for their help in completing this thesis. First of all I would like to thank my supervisors Julie Logan, Kirstin Edwards and Christine McCartney for their help, support and probably most importantly patience. I would also like to thank Saheer Gharbia, mostly for her scientific input but also for providing a number of well-timed pushes throughout this project. For the bioinformatics work I would like to thank Antony Underwood and Raju Misra for support and training, they were very important in developing my bioinformatics skills to the level they are now at and particular thanks go to Nadia Ahmod who provided the initial scripts that formed the basis of a large part of this project as well as helping develop my skills. Unfortunately she passed away after a long illness before this project could be completed. It is a great shame that no one else will be able to receive the benefit of her wisdom in the future. For the laboratory work I would like to thank Chloe Bishop, Sally Langham, Dunstan Rajendram, the staff of NCTC for training and assistance as well as the rest of the staff of DBHT. Last but not least I would like to thank my wife Megan, mainly for putting up with me throughout the PhD process and also for providing a more or less never-ending supply of coffee and toast, without this support, it is unlikely this thesis would ever have been completed.

Abstract

Sequence typing is a rapidly evolving field and offers improved analysis into the genetic background and lineages of organisms compared to serological or DNA banding pattern based analysis. However, the resolution of molecular typing schemes varies between organisms and often loci used in sequence typing lack discriminatory power and give limited information into the evolution of the organism. This is particularly true of group B streptococci (GBS), where the same sequence types appear worldwide, which is unlikely for such a pathogen.

This project aimed to develop a two component pathogen profiling approach which accurately reflected the phylogeny of GBS isolates, using elements of the core genome and elements of the variable genome. To address this a bioinformatic approach which selected loci for sequence typing based on predicting genes which evolve in the same manner as the average for the core genomes was adapted from a previously applied study for designing genus level sequence typing schemes. This informed the selection of candidate loci which were then experimentally verified, using a collection of 135 GBS clinical isolates. It was demonstrated that it was possible to obtain greater resolution and accuracy using only three unique genes that are intelligently selected, rather than using seven known housekeeping genes that are selected at random.

Sources of hyper variability within the genome, in particular the presence of mononucleotide repeats (MNR) were investigated in non-coding DNA. It was postulated that these regions of DNA are more prone to mutation due to the lack of selective pressures, the presence of MNR repeats make these regions more unstable during replication and that in core genes these regions may be involved in genomic regulation by slipped strand mispairing. Results did confirm that non-coding DNA containing MNR repeats were more variable than DNA without them but these did not match the discriminatory power of MLST typing or the new three gene typing scheme. However, it was observed that one MNR tract was an insertion site for one of two insertion sequences and that typing using the presence/absence of these insertion sequences further enhanced

discriminatory power in addition to the 3 gene scheme and may yet prove to be an indicator of virulence in clinical GBS isolates.

As well as demonstrating that sequence typing can be more informative if sequence typing markers are intelligently selected, this project also showed the importance of developing computational methods to analyse pathogen genome sequences that are being released in ever increasing numbers thanks to new and constantly improving technologies. For example, methods used here to determine the core and pan-genome of any given set of genomes are becoming increasingly important to the study of pathogen evolution, virulence and population dynamics.

Contents

List of Abbreviations

AAP	American Academy of Pediatrics
ABCs	Active bacterial core surveillance
ABI	Applied Biosystems
Abs	Absolute subtraction
ACOG	American Congress of Obstetricians and Gynecologists
ANI	Average nucleotide identity
ATP	Adenosine triphosphate
BLAST	Basic local alignment search tool
bp	Base pairs
CAMP	Christie Atkins Munch-Petersen
CDC	Centre for Disease Control
CI	Confidence interval
COG	Clusters of orthologous groups
CPS	capsular polysaccharide
dATP	Deoxyadenosine triphosphate
dCTP	Deoxycytidine triphosphate
dGTP	Deoxyguanosine triphosphate
dN/dS	Synonymous/non-synonymous

DNA	Deoxyribose nucleic acid
dNTP	Deoxyribonucleotide triphosphate
dTTP	Deoxythymidine triphosphate
EOD	Early onset disease
FASTA	Fast all
FRET	Förster resonance energy transfer
GAS	Group A Streptococcus
GBS	Group B Streptococcus
GFS	Group F Streptococcus
GGS	Group G Streptococcus
GTR	General time reversible
HPA	Health Protection Agency
IUPAC	International Union of Pure and Applied Chemistry
JCVI	J. Craig Venter Institute
LOD	Late onset disease
MALDI-TOF MS	Matrix assisted lazer deabsorbtion/ionisation time of flight mass spectrometry
MCMC	Markov chain Monte Carlo
MGEs	Mobile genetic elements
MLST	Multi locus sequence typing

MNRs	Mononucleotide repeats
NCBI	National Centre for Biotechnology Information
ORF's	Open reading frames
PAUP*	Phylogenetic analysis using parsimony (and other methods)
PBS	Phosphate buffered saline
PCR	Polymerase chain reaction
PERL	Practical extraction and report language
PFGE	Pulse field gel electrophoresis
PPi	Inorganic pyrophosphate
QV	Quality value
RAPD	Random amplified polymorphic DNA
RNA	Ribosomal ribonucleic acid
rpm	Revolutions per minute
rRNA	Ribonucleic acid
SAP	Shrimp alkaline phosphotase
SELDI-TOF MS	Surface-enhanced laser desorption/ionization time of flight mass spectrometry
SMRT	Single molecule real time
SNP	Single nucleotide polymorphism
SSRs	Short sequence repeats

ST	Sequence type
START2	Sequence Type Analysis and Recombinational Tests Version 2
TBE	Tris/Borate/EDTA
ti/tv	Transition/transversion
UK	United Kingdom
UPGMA	Unweighted pair group method with arithmetic mean
USA	United States of America
VF	Virulence factor
VFDB	Virulence factors database
wgs	Whole genome shotgun

List of Figures

List of Tables

Chapter 1

Introduction

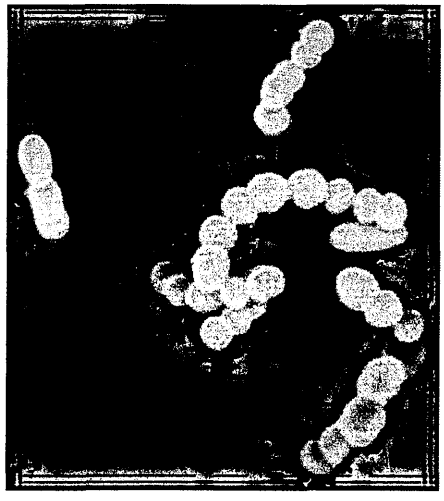
1.0 Introduction

1.1 Streptococcus agalactiae

1.1.1 Biology

The genus *Streptococcus* are Gram positive bacteria with spherical or ovoid cells arranged in chains or pairs (figure 1.1). All species are non-motile, non-spore forming and have complex nutritional requirements. They are obligate parasites of mucosal membranes. Some *Streptococcus* species are members of the commensal microflora and others are highly pathogenic. *Streptococcus agalactiae* (group B streptococcus or GBS) is a combination of the two, it can colonise the gastrointestinal and genitourinary tract without causing symptoms but if it becomes established in a normally non-sterile site it can cause severe invasive disease (112). It was first isolated from cattle as the cause of bovine mastitis with other pathogenic streptococci and first differentiated into group B streptococci in 1934 by Rebecca Lancefield (126). GBS emerged as a serious human pathogen in the 1970s and transfer of virulence genes from group A streptococcus has been presented as one possible explanation for its sudden emergence

Figure 1.1: Scanning electron micrograph of the chains of GBS bacterium



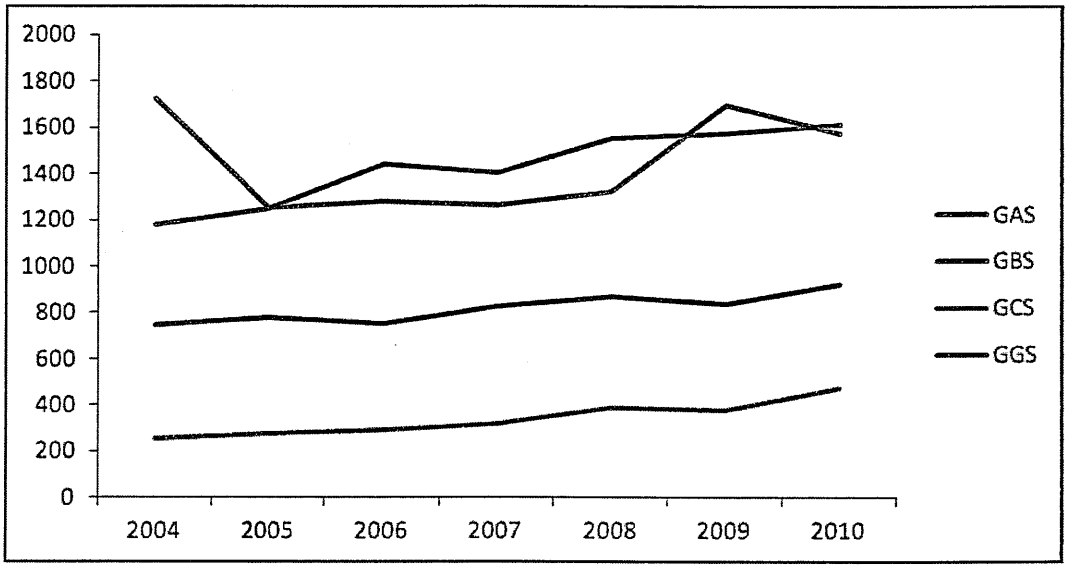
Source: Craig Rubens, University of Washington

1.1.2 Epidemiology, Pathogenesis and Clinical Disease

1.1.2.1 Trends in GBS Incidence in the UK

Incidence of bacteraemia caused by GBS in the UK increased substantially between 2005 to 2006 from 1249 to 1442 (15% increase), after a slight decrease in incidence between 2006-2007 (1442 to 1403). A further increase was observed between 2007-2008 from 1403 to 1550 (10% increase). 2008-2009 also showed an increase from 1550 to 1571 (1.3% increase) and the latest year for which figures are available 2009-2010 showed an increase from 1571 to 1610 (2.5% increase) making the total increase from 2004-2010 of 27% meaning the disease burden caused by GBS is on a long upward trend in the UK. As shown in Figure 1.2 the increased incidence of bacteraemia caused by GBS combined with a decrease in incidence of bacteraemia caused by group A streptococci (GAS) between 2004 and 2010 (1754 to 1574) make GBS bacteraemia the current leading cause of pyogenic streptococcal bacteraemia in England, Wales and Northern Ireland (85).

Figure 1.2: Bacteraemia caused by pyogenic streptococci (group A streptococci, GAS, group B streptococci, GBS, group C streptococci, GCS and group G streptococci, GGS) in England, Wales and Northern Ireland 2004-2008.



Source: Pyogenic and non-pyogenic streptococcal bacteraemia, England, Wales and Northern Ireland: 2010, Health Protection Agency

The overall rate of GBS bacteraemia in 2010 for England, Wales and Northern Ireland was 2.8 per 100,000 population the same rates are observed in Wales and England although Northern Ireland has a much higher level of incidence (3.4/100,000 population). Within England, rates vary considerably by region with the lowest rates being observed in the North East and South East (both 2.0/100,000) and the highest rates outside of Northern Ireland being the West Midlands (3.3/100,000). Rates of GBS bacteraemia were highly concentrated in infants, 71 per 100,000 population <1y, with higher rates in males than females with the exception of 15-44 year olds (2.0 and 0.6 in females and males respectively). Untreated, mortality rates are high (50%) and many survivors showed permanent neurological sequelae. Use of intrapartum antibiotic prophylaxis administered during labour reduces mortality rates to ~15% (209).

1.1.2.2 Epidemiology of Early Onset and Late Onset Disease

In the UK early onset disease (EOD) is more common than late onset disease (LOD) with 0.39/1000 live births opposed to 0.27/1000 live births (85) (Table 1.1). However, the incidence of LOD is more commonly associated with the onset of meningitis with 43% of cases presenting with meningitis compared to 11% for cases of EOD. Conversely, septicaemia is more prevalent in EOD (63%) than LOD (41%) (128). Of the number of cases developing meningitis, half develop long term disability including mental retardation and loss of vision (13) and since meningitis is more common in LOD the majority of long-term disability is caused by LOD even accounting for the lower number of cases, whilst EOD has a higher case fatality rate of 10% compared to 8% for LOD (86). However, it is worth noting that these rates may be overestimated since the UK has no universal screening system and less severe cases of EOD or LOD may be missed (128)

Table 1.1: The onset of GBS bacteraemia split according to age at onset and the 95% confidence interval (CI)

	Number	Rate/1000 Live Births	95% CI
Total Cases (0-90 Days)	506	0.69	0.63-0.76
Early Onset (0-6 Days)	302	0.41	0.37-0.46
Late Onset (7-90 Days)	204	0.28	0.24-0.32

Source: Pyogenic and non-pyogenic streptococcal bacteraemia, England, Wales and Northern Ireland: 2010, Health Protection Agency

1.1.2.3 Pathogenesis of Early Onset Disease

Research into GBS has focused on the organism's ability to cause invasive disease in neonates since this is the most common form of the disease. Disease in neonates can be split into two forms, early onset (<7 days) and late onset (7-90 days) which is less common. Pathogenesis of early onset invasive neonatal disease occurs first through asymptomatic colonisation of the genitourinary tract of a pregnant mother, shown in 38% of adult women (64) followed by transmission to the neonate which occurs in 50-70% of births (167). Acquisition by the infant may occur by one of two mechanisms: 1) exposure *in utero* after ascending infection of the placental membranes and amniotic fluid; or 2) transmission through passage of an infected birth canal. In the first stage of infection GBS enters the lung causing infected neonates to present with respiratory symptoms. Autopsies of early onset fatalities show 80% have histological evidence of lobar or multilobar pneumonia. From the lung GBS traverses three host barriers, the alveolar epithelium, the pulmonary interstitium and the pulmonary endothelium, which damages lung tissue and allows GBS entry to the blood stream. The host immune response to GBS in the blood causes sepsis syndrome and in some cases septic shock, which is clinically indistinguishable from Gram-negative endotoxemia. From the blood, GBS can cross the blood brain barrier resulting in meningitic incidence (167).

1.1.2.4 Pathogenesis of Late Onset Disease

Whereas the mode of infection for early onset GBS infection is relatively well understood, the mode of infection for late onset disease in less well understood. A number of potential modes of infection have been postulated including acquisition of the organism vertically during vaginal delivery, or postnatally from maternal/carer contact, infected breastmilk (168) or nosocomial sources (177). As with early onset disease premature delivery is still considered a risk factor (187). It may be possible to identify the most likely level of transmission in individual cases by considering the method of delivery (vaginal or caesarean), the infection status of the mother or method of feeding i.e. breast or formula (45).

After infection, pathogenesis is broadly similar to early onset disease except that meningitis is more commonly associated with late onset disease. Either because the host immune system is more developed and therefore more likely to trigger inflammation around the brain or the fact that early onset disease shows more severe pneumonia and sepsis syndrome causing fatality before meningitis sets in (167).

1.1.2.5 Adult Infection

GBS has also been shown to cause disease in adult patients, the most common presentation being skin and soft tissue infections but urinary tract infections (239), meningitis and bacterial sepsis have been documented. Cases in adult populations more frequently affect the elderly and immunocompromised but cases in patients with no known risk factors have been described. One study of GBS soft tissue infections showed that as high as 24% of patients had no obvious underlying conditions, mortality rates reached 7% and 11% of patients lost limbs through infection (130). The pathogenesis of GBS in adult infections usually relies on failures in the host immune system, for example one risk factor for GBS soft tissue infection is cutaneous ulceration demonstrating that disruption to the skin allows infection to become established and in some cases proceed into the blood stream (130). It is also possible that a combination of physical injury,

pre-existing conditions and mixed infections can contribute to cause skin infections, one example being Necrotizing Fasciitis caused by *Streptococcus agalactiae*, *Arcanobacterium haemolyticum*, and *Finegoldia magna* in a dog-bitten patient with diabetes (131). In patients with no known risk factors novel modes of transmission have been observed for example GBS entering the blood stream through lesions in the oral cavity caused by tooth brushing (71).

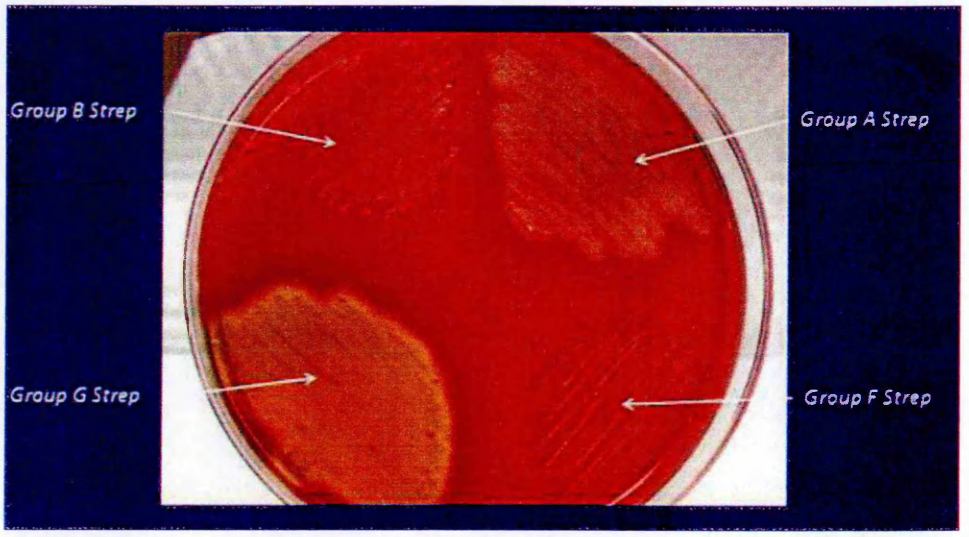
Pregnant women as well as being at risk of passing on the infection to neonates can also develop symptoms from infection including fever, premature labour and symptoms associated with a urinary tract infection and these symptoms can be used as risk factors of transmission to the neonate (128).

1.1.3 Detection and Identification

1.1.3.1 Culture

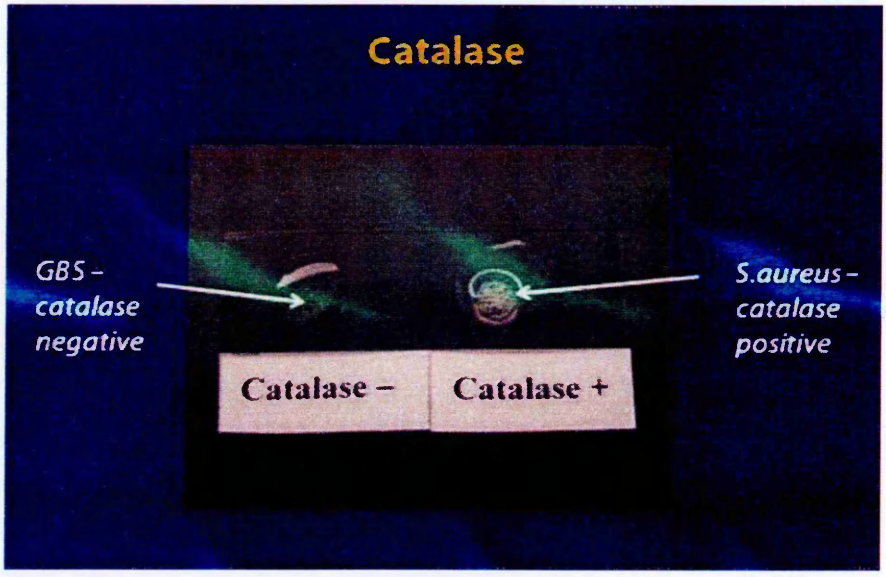
Standard laboratory identification of GBS is performed by culture, cells are grown on blood agar and GBS identification is positive if: β-haemolysis occurs on a Columbia agar plate containing 5% horse blood (figure 1.3), Gram staining shows Gram-positive cocci in pairs or short chains, there is a negative reaction with catalase reagent (figure 1.4), and Lancefield grouping with type B antisera is shown (207). Additionally the Christie Atkins Munch-Petersen (CAMP) test can identify GBS strains by showing lysis of sheep or ox cells when grown in the proximity of *Staphylococcus aureus* strains under anaerobic conditions (figure 1.5). Detection in a clinical setting is commonly performed by selective culturing on enriched medium that will allow preferential growth of GBS when inoculated with samples taken from vaginal or rectal swabs. For example the StrepB carrot broth (Hardy Diagnostics, Santa Maria CA, USA) is a pigmented enrichment broth which positively identifies GBS if there is a positive colour change (figure 1.6).

Figure 1.3: β-haemolytic activity of GAS, GBS, GFS and GGS.



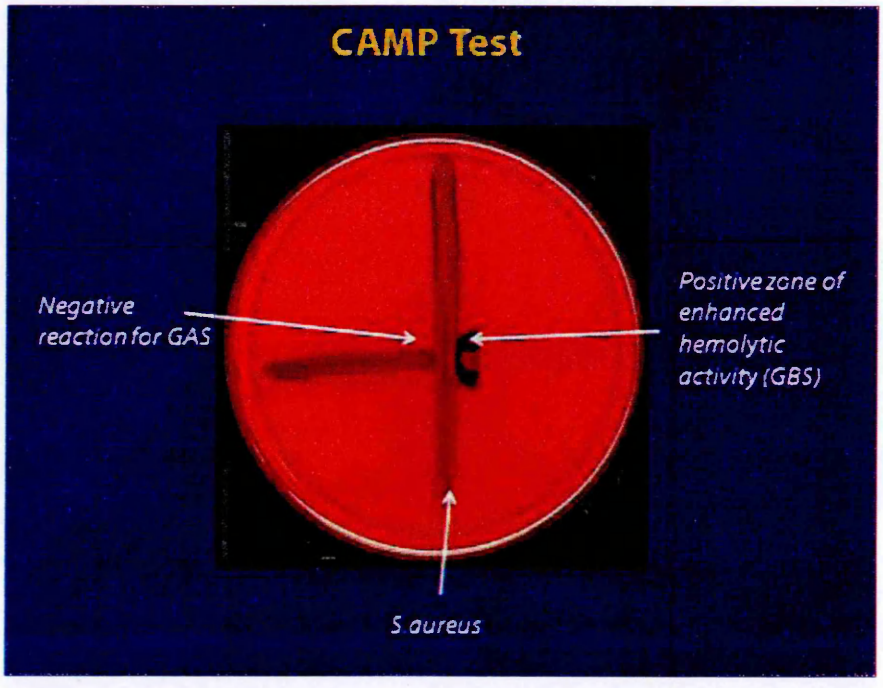
Source: CDC Laboratory Slide Set

Figure 1.4: Negative catalase reagent test



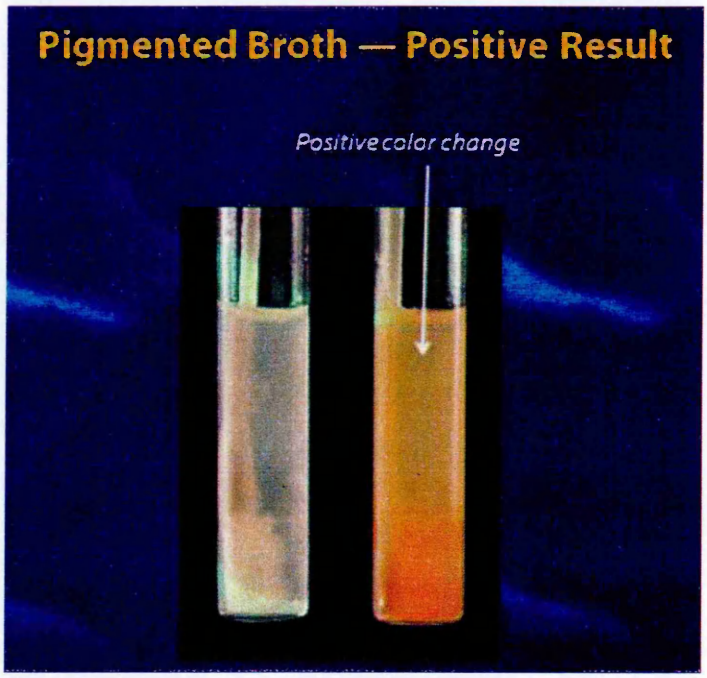
Source: CDC Laboratory Slide Set

Figure 1.5: Positive CAMP test



Source: CDC Laboratory Slide Set

Figure 1.6: Positive GBS test using the Strep B carrot broth

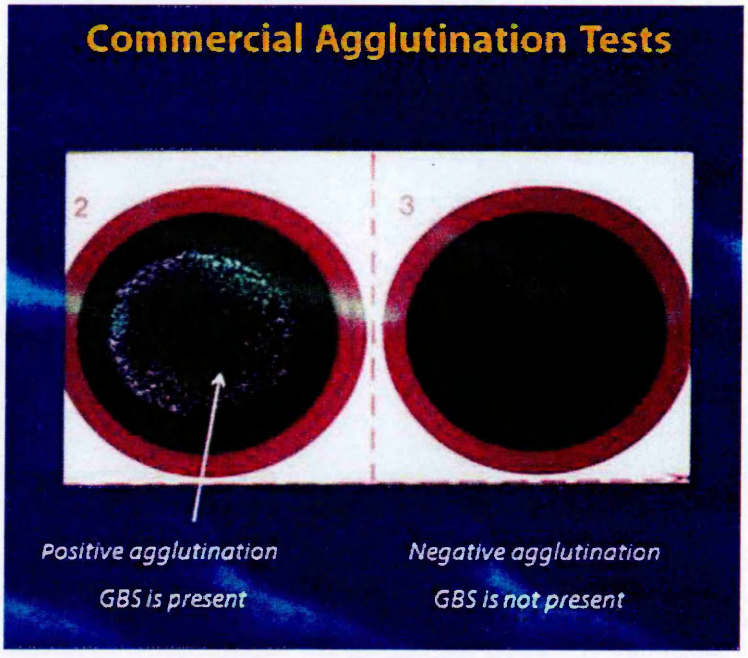


Source: CDC Laboratory Slide Set

1.1.3.2 Immunological Detection

GBS carries the Lancefield group B antigen and detection methods using latex agglutination and immunoassays for the detection of this antigen have been developed (215) (figure 1.7). The HPA *Streptococcus* Reference Laboratory uses the Streptex kit (Remel, Lenexa KS, USA) to confirm the identity of cultures. This assay requires culturing of GBS prior to performing the test and can therefore only be used to confirm culture-based identification. Two early studies on similar immunological assays have been carried out to assess their sensitivity and specificity when applied directly to clinical samples which showed that when comparing selective broth culturing to immunoassays for identification of GBS direct from vagino-rectal swabs, sensitivity of between only 4%-37% was observed (8,257). However, the prevailing view is that latex agglutination assays give a high level of false positive and false negative results (51). Therefore these tests can only be used to confirm culture identification or identify heavily colonised patients (180).

Figure 1.7: Positive and Negative latex agglutination tests for GBS



Source: CDC Laboratory Slide Set

1.1.3.3 Molecular Detection

The Polymerase Chain reaction (PCR) was developed by Kary Mullis in 1983 and is a molecular biology technique to amplify from a small number of copies of a piece of DNA, generating thousands to millions of copies of a particular DNA sequence selected by the use of target specific primers and thermostable enzymes.

PCR has been used for the identification of GBS where loci unique to GBS have been amplified by conventional or real-time PCR (54). There is a wide array of assays for clinical diagnosis and research which amplify different genomic targets e.g the genes *cfb*, encoding CAMP factor and *scpB* which encodes segregation and condensation protein B (110,190).

Diagnostic assays are moving towards using real-time PCR (39,54,106,166,190,227,238) in which amplification is measured in real-time by using sequence-specific probes or dyes that fluoresce when they bind to or intercalate with double stranded DNA. The interaction of the probes or

fluorescent dyes with double stranded DNA is measured at each cycle and used to calculate the amount of DNA in the reaction. Real-time PCR has several advantages over conventional PCR, it is faster, it can be carried out in a closed tube system preventing carry over contamination and results are immediately accessible since no post amplification visualisation of product is required.

Identification using a PCR assay offers several advantages over culture. DNA can be isolated directly from vaginal and/or rectal swabs of pregnant women eliminating the need for time consuming culture (238). Using PCR can determine the GBS colonisation status of mothers and identify infection in neonates faster than culture based methods, and it could even be used as a point of care test speeding up detection further. PCR methods have also been shown to be superior in identifying colonisation than culture. For example, Natarajan et al. showed that a real-time PCR assay for the *cfb* gene had a sensitivity of 90% and detected GBS carriage in 51% of women tested compared to culture which only identified colonisation in 17% of women (166). However, PCR may overestimate the rates of colonisation because PCR cannot distinguish between live and dead bacteria and since PCR is more sensitive that culture, low level colonisation that may not be able to cause disease can also be identified, leading to overuse of prophylactic antibiotics.

Finally, generic real-time PCR assays targeting the 16s rRNA fragments have been developed to distinguish bacterial septicaemic disease from other causes of neonatal illness such as asphyxia or complications of prematurity. These have been used with varying success in the analysis of whole blood for neonatal sepsis, specificity is generally high but sensitivity can be as low as 40% (106,107).

1.1.3.4 Novel Methods

Although not a method used for diagnosing GBS infection, surface-enhanced-laser-desorption-ionization time of flight mass spectrometry (SELDI-TOF MS) has been successfully used to identify four amniotic fluid proteomic biomarkers, namely human neutrophil defensins two and one and calgranulins C and A. These have been shown to be strongly predictive for neonatal sepsis and neonates born to women, with three out of four of these biomarkers shown to have an association with increased incidence of sepsis (25,26). Using SELDI-TOF MS in this manner has advantages and disadvantages. It's advantages include being a rapid catch all for neonatal sepsis allowing treatment to be rapidly applied. Conversely, since this method is diagnosing sepsis rather than a specific infection no causative agent is known meaning organism specific qualities such as antibiotic resistance could be missed.

1.1.4 Treatment and Prevention/Diagnosis

1.1.4.1 Treatment

If GBS is detected prior to birth or risk factors such as prolonged rupture of membranes or fever in labour may indicate infection, treatment is intrapartum antibiotics given to the mother during labour. If transmission to the neonate is not prevented and the neonate shows signs of infection intravenous antibiotics and intensive care treatment are used. The standard antibiotic used for treatment for GBS infection is penicillin, with clindamycin or erythromycin used in patients allergic to penicillin. Figures for the UK from 2003 to 2010 show an increase in resistance to erythromycin from 7% to 15%, while clindamycin levels fluctuate around 8-9% when figures where last available (85). There are two main mechanisms conferring resistance to these antibiotics. The first is erythromycin ribosomal methylase (mediated by *ermB, ermA, ermTR* or *ermC* gene variants) which confers cross-resistance to macrolides, lincosamides and streptogramin B (77). The second is a macrolide efflux pump (mediated by *mef*) which confers resistance to 14 and 15 member macrolides only (27).

It has been suggested that the use of intrapartum antibiotics could lead to an increase in the prevalence of resistant organisms and cause adverse effects in neonates. For example, Stoll et al. showed that increased antibiotic use in neonates has led to decreased rates of early onset GBS infection but also led to a rise in early onset *E. coli* sepsis leaving overall rates of early onset sepsis in neonates stable (226). However, this conclusion is far from the consensus. Cousens et al. showed that intrapartum antibiotic use has decreased the level of complications from preterm delivery, reduced rates of neonatal infections and reduced overall neonatal mortality (36) and Balter et al. showed that neonates born to mothers treated with intrapartum antibiotics were no more or less likely to undergo invasive procedures or to receive further antibiotic treatments (10).

1.1.4.2 Prevention/Diagnosis

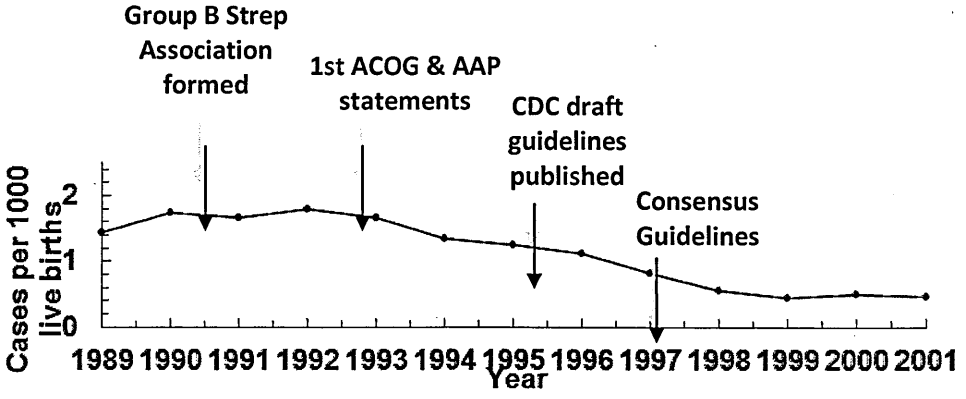
1.1.4.2.1 Risk Factor Based Screening

In the UK national policy dictates prevention of GBS infection in neonates by a risk factor-based screening approach for infection in pregnant women using five risk factors, preterm labour, prolonged rupture of the membranes, fever in labour, GBS bacteriuria detected during the current pregnancy and previous baby affected by GBS infection (196), with the first three risk factors considered to be the most important (128). It is predicted that this risk factor based approach is 67% effective at preventing early onset GBS disease with a 17% false positive rate (128). However, despite national guidelines being issued in 2003 they have not yet been implemented nationally and only 34 of 227 maternity units could be shown to be following national guidelines exactly (38).

1.1.4.2.2 Culture Screening

In the US, Centers for Disease Control and Prevention (CDC) guidelines recommend culture screening all pregnant women between 35-37 weeks gestation (207). Culture methods identify more cases of GBS colonisation that risk factor screening with 90% GBS positive mothers identified by culture compared to 67% by risk factor screening alone (128). There are weaknesses in this approach, costs are high, GBS colonisation can be transient so results at the time of testing may not reflect colonisation status during labour and rupture of membranes caused by GBS can trigger premature labour which can put birth before the 35 week mark when culture tests would be performed. However, culture based methods are still an improvement on a risk factor based approach since it can be shown on a national level that GBS incidence rates have fallen in the USA since the introduction of CDC guidelines recommending the introduction of culture screening between 35-37 weeks gestation. This is shown in figure 1.8 and can be compared to the increasing levels of GBS infection in the UK shown in figure 1.2.

Figure 1.8: The decrease in GBS incidence in the USA after introduction of CDC guidelines



Source: CDC Active Bacterial Core surveillance (ABCs) program

1.1.4.2.3 PCR versus Immunological screening

Although there are no national guidelines that use either PCR or immunological methods as a standard method for GBS screening there have been a number of studies to suggest these methods may be more effective than culture. Firstly, PCR based assays are faster than culture and can be performed in either a laboratory setting with appropriate equipment or theoretically in a ward or field setting using portable real-time PCR system such as the RAZOR EX (Idaho Technology Inc, Salt Lake City UT, USA) which would overcome some of the drawbacks of culture-based screening.

The sensitivity and specificity of PCR based assays has been shown to be either comparable to culture screening, for example the *RiboSEQ* GBS test was shown to be 96.4% sensitive and 95.8% specific compared to culture screening (252) or superior to culture. For example, Rallu et al showed that PCR assays for the *scpB* and *cfb* genes had sensitivity and specificity rates of 99.6% and 100% and 75.3% and 100% respectively, compared to a sensitivity of 42.3% and specificity of 100% for the standard CDC culture method (190). This study also compared the PathoDX (remel) culture identification method and found an improvement on the CDC culture screening but was not as sensitive or specific as either of the PCR methods with a sensitivity of 57.3% and a specificity of 99.5%. However, each of these three methods were carried out on cultured organisms, not directly from patient swabs. Another study found that real-time PCR screening was able to identify three times more cases of GBS colonisation in pregnant women with 51% of women assessed by real-time PCR opposed to 17% assessed by culture being identified as colonised by GBS as the time of delivery. This study showed that real-time PCR would be a useful clinical tool in the management of infants potentially at risk of invasive GBS infection. (166).

1.1.5 Vaccination

In theory vaccination would eliminate infection and the need for screening programs. However, since in practice uptake of vaccines is never 100% and herd immunity is unlikely to be applicable to an organism that is for the most part a commensal organism even vaccination would not completely remove the burden of GBS disease. Antibody levels in the neonate remain constant for up to three months after birth and therefore would also protect against late onset disease. It would remove the need for antibiotics preventing an increased risk of allergic reactions later in life and prevent the removal of penicillin sensitive organisms allowing increased growth of antibiotic resistant bacteria (128). Natural immunity to GBS can be conferred by transmission of capsular polysaccharide (CPS) serotype specific IgG antibodies across the placenta (9). Because of this a number of vaccines are composed of capsular polysaccharides conjugated to a protein which improve the carbohydrates immunogenicity (76). For example, a tetravalent vaccine containing tetanus toxid conjugated to CPS Ia, Ib, II and III is predicted to prevent 90% of cases of invasive neonatal disease (176). However CPS vaccines would require updating as new serotypes emerge in human disease. Sequencing of GBS genomes has also allowed reverse vaccinology (100,192) research to develop a vaccine based on universal cell surface protein antigens which may overcome this problem (147,211). Other advantages include the reduced time of vaccine development, theoretically reducing development time from 5-15 years to 1-2 years (100). Other problems with vaccine development are political and economic, such as reluctance to develop products for use during pregnancy for the fear of teratogenicity despite at least one vaccine being proved non-teratogenic in rabbits (174) and perceived limited returns due to scepticism surrounding vaccines in general (128).

1.2 Genomics of GBS

1.2.1 Typing Methods

1.2.1.1 Serotyping

GBS can be classified by serotyping (126) on the basis of ten type specific capsular polysaccharides (CPS) IA, IB, II, III, IV, V, VI, VII, VII and the recently proposed IX (216). They can be further classified according to three cell surface localised protein antigens C, X and R. A recent meta-analysis has shown that the most common serotypes worldwide are serotypes Ia, III and V which account for 72% of GBS isolates (95). Serotyping is most commonly performed by latex agglutination using antibodies specific to each CPS and protein antigen (215). Serotyping via latex agglutination has been used since it is a rapid, simple method for typing GBS. Serotyping may have been superseded by molecular methods that better differentiate GBS isolates but serotyping is still used for historical reasons since this method has been used since the discovery of GBS in 1934 and by maintaining this method all GBS clinical isolates that have been serotyped are comparable. Secondly vaccine development is looking at CPS variants as vaccine candidates so the serotypes in the population must be monitored to select the most appropriate candidate(s) for vaccination (20,251). However, there are limitations to serotyping including non-typeable isolates, one study showed 12% of isolates were nontypeable by latex agglutination (116) and some serotypes show cross-reactivity resulting in a non-conclusive result. Finally it has been shown that there is significant genetic variation between isolates of the same serotype as shown by MLST and PFGE (58,103,191), meaning that serotypes do not reflect genetic diversity.

1.2.1.2 Molecular Methods

1.2.1.2.1 Random Amplification of Polymorphic DNA

Random Amplification of Polymorphic DNA (RAPD) was originally developed in 1990 by Williams et al. (255) and is a method to construct genetic maps without prior knowledge of DNA sequence. In this method random 8-12 nucleotide primers are used to amplify regions of chromosomal DNA, resulting in the amplification of a number of different size PCR products which are visualised by agarose gel electrophoresis. Polymorphisms are identified by the presence/absence of each band and so RAPD can be used to place isolates into clusters. Using RAPD, Duarte et al. showed relationships between GBS isolates from cows of the same herd (48) and Martinez et al. showed that human isolates clustered separately from bovine isolates (154). It can also be used to identify regions of the genome that are linked to pathogenic traits, since missing or additional bands can indicate insertions and deletions that render priming sites too distant to allow amplification, or insertions/deletions that change the size of a DNA segment without preventing its amplification (255). For example, van der Mee-Marquet et al. used RAPD to identify prophagic DNA fragments associated with virulent GBS strains (244).

RAPD has largely been replaced by other profiling methods such as PFGE and MLST for typing of GBS clinical isolates. However a number of research groups still use RAPD. The most common use of RAPD is in epidemiological studies measuring transmission of isolates between defined patient groups and it has been used to show direct transmission or lack of direct transmission between patients (52,53,82,243). It has also been used for confirmation of results generated using other profiling methods (151).

1.2.1.2.2 Pulsed Field Gel Electrophoresis

Pulsed Field Gel Electrophoresis (PFGE) was developed as a method to view large DNA molecules of up to 2000kb via gel electrophoresis and was originally used to separate intact chromosomes of *Saccharomyces cerevisiae*. The technique uses alternatively pulsed, perpendicularly oriented electrical fields at least one of which is inhomogeneous. The duration of the electrical pulses is varied from 1 second to 90 seconds to achieve optimal separations of DNA molecules ranging from 30 to 2000 kb in size (210). For fingerprinting of bacterial genomes the bacterial chromosome is fragmented using a restriction enzyme that cuts the genome at a small number of restriction sites (158). Following digestion, DNA is subjected to PFGE as described above and the position of fragments on a gel, which indicates the size of the fragments, is used to cluster bacterial isolates. PFGE has been used in typing GBS isolates to show that isolates within the same serotype show considerable genetic variation (58), that particular fragments detected can be linked with virulent isolates from CSF samples (194) and has clustered 35% of macrolide-resistant isolates into one PFGE type (44). PGFE is also useful for identifying chromosomal insertions or deletions, for example Bohnsack et al. used PFGE to show the presence of two GBSi1 introns in serotype III GBS isolates by selecting relevant bands in the PGFE profiles where these introns may exist on the basis of their size and then confirmed their presence by subtractive hybridization (18). Martins et al. have also used PFGE to aid in the identification of rare capsule switching events in GBS (155).

1.2.1.2.3 MLST

MLST is a sequence based molecular typing system in which a number of conserved housekeeping genes, usually seven, are sequenced and used to place strains into clusters known as sequence types. MLST was originally developed by Maiden et al. for *Neisseria meningitidis* (146) and to date has been used on more than 80 different species.

In the GBS MLST scheme (103) fragments of 7 housekeeping genes involved in intermediary metabolism are sequenced (*adhP*, *atr*, *glcK*, *glnA*, *pheS*, *sdhA* and *tkt*). These loci were selected for their presence in all GBS strains (as evidenced by their use in MLEE typing systems) and their chromosomal location, since loci too close together cannot be used because recombination events may exaggerate evolutionary relationships (103). The sequenced loci are assigned an allele type on the basis of similarity to previously identified alleles using the online MLST database (101). Additionally in this method allelic sequences can be concatenated and used for phylogenetic analysis to study the relationships between strains. However, using concatenated sequences of the MLST loci to infer phylogeny has been shown to poorly reflect whole genome phylogeny in *E. coli* (119).

The advantages of using sequence data is that it is comparable between laboratories worldwide, whereas gel based methods such as RAPD and PFGE often show inter-laboratory variation. MLST also gives a larger number of unique sequence types when compared to either RAPD or PFGE. However, there are limitations to its discriminatory power as it clusters large number of isolates together. For example, the initial work when the GBS MLST scheme was developed examined 152 isolates and placed them into only 29 sequence types with the majority of isolates (101/152) placed into only 4 sequence types (103). Further work using 338 strains gave an additional 29 sequence types with the top 6 sequence types (all containing > 25 isolates) containing 281/388 (72%) of all isolates (149).

1.2.1.2.4 Molecular Serotyping

Molecular serotyping is a sequence-based method which aims to reproduce conventional serotyping by sequencing capsular polysaccharide (*cps*) gene clusters (115) as well as the genes encoding the cell surface antigens C, R and X (114). This method has been used successfully to serotype isolates that are non-typeable by immunological serotyping and could assign molecular serotype to 98.5% of isolates (116). However, the molecular serotypes assigned may not agree with conventional serotyping as demonstrated by Manning et al. who showed that molecular serotypes Ia and III could show either conventional serotype (149). Chaffin et al. demonstrated that this was due to the *cpsH* gene controlling CPS expression independent of the rest of the CPS operon (30). Molecular serotyping has been applied to a number of other organisms including *E. coli* (50) and *Streptococcus pneumoniae* (7).

1.2.1.2.5 Three Set Genotyping

Three set genotyping, proposed by Kong et al. in 2003 (117), is an extension of the molecular serotyping system. As well as sequencing *cps* genes and genes encoding cell surface protein antigens, a set of 5 mobile genetic elements (MGE's) are sequenced if they are present. Sequence type is assigned on the basis of allele sequence and presence/absence of the five MGE's (117). As with MLST large numbers of isolates are placed into a limited number of groups. In one study, the three set genotyping system places 83 isolates into 27 genotypes with the top four genotypes covering over 60% of tested isolates compared to 24 MLST types with 60% of isolates being found in the top four genotypes (228).

1.2.1.2.6 Sequence Typing using Non-Coding DNA

Various sequence typing methods have been developed using markers from non-coding regions of the genome. These include 16S–23S rRNA gene internal transcribed spacer (ITS) (66,199), Multi Spacer Typing (MST) (47) and Multiple Loci VNTR Analysis (MLVA) (144,240,245). The assumption behind sequence typing methods from non-coding regions is that non-coding DNA is theoretically unaffected by selective pressure and therefore can be more variable and will provide more discriminatory strain typing systems (47).

ITS typing was the first one developed and is useful because like the 16S rRNA gene it is common to all bacteria with the exception of *Rickettsiales* (138). It is also present in multiple copies per genome (21). Additionally, it is more variable than the 16s rRNA gene making this method more discriminatory.

As well as single non-coding regions for typing, multiple regions can be selected for various organisms to create Multi Spacer Typing (MST). MST loci can be selected either because of the presence of repeat regions which should also be more variable due to replication errors (74) or non-coding sequences which show significant variation between different bacterial strains. For example MST based on six intergenic spacers divided 36 *Y. pestis* strains from three biovars from dental pulp of patients deceased from plague in the second and third pandemics into 19 sequence types (47). MST has also been applied to *Rickettsia conorii* (63), *Rickettsia prowazekii* (260), *Rickettsia sibirica* (62), *Coxiella burnetii* (69), *Bartonella henselae* (134,136), *Bartonella quintana* (61) and *Tropheryma whipplei* (137) and when MST schemes are compared to MLST schemes MST is shown to be more discriminatory (63,136).

Multiple Loci VNTR Analysis (MLVA) is used to type organisms based on the number of tandem repeats found at any given locus that varies in copy number. These repeats and are dispersed widely in both human and bacterial genomes (144,240,245). In bacterial genomes, VNTR loci are found in non-coding regions as well as in genes and these non-coding VNTRs make good targets for strain typing because of their rapid evolution (63,170,240) and therefore MLVA tends to be

more discriminatory than other methods. For some species which show high levels of homology such as *Francisella tularensis* (57,99), *Bacillus anthracis* (89,111,140), *Yersinia pestis* (113), and *Mycobacterium tuberculosis* (129) MVLA typing is considered the gold standard. MLVA has also been applied to other important human pathogens with varying levels of success such as meticillin-resistant *Staphylococcus aureus* strains (230), *Burkholderia pseudomallei* (236), and *Clostridium difficile* (241) and even GBS (186).

The rapidly evolving nature of non-coding loci, particularly repeat regions the main weakness of typing using non-coding DNA. For example MLVA typing of *Mycobacterium leprae*, has shown variation in the VNTR pattern between isolates of *M. leprae* biopsies from the same patient (162). Although this effect may be dependent on the species or even the specific target as it has also been shown that MLVA for *Enterococcus faecium* is less discriminatory than PFGE and MLST (253).

1.2.2 Whole Genome Analysis

The ability to sequence whole genomes has revolutionised microbiology, allowing characterisation of bacteria at the genus, species and strain level (231). At the outset of this study there were 3 fully sequenced GBS genomes 2603V/R, NEM316 and A909 and 5 partially sequenced "shotgun" genomes 18RS21, 515, CJB111, COH1 and H36B available. Currently, there are an additional 273 partially sequenced genomes that have recently been published by the JCVI.

Sequencing of single genomes allows general features to be characterised. For example, the G+C content of GBS strains NEM316 and 2603V/R is 35.6% and 35.7% respectively (68,232). Open reading frames (ORF's) in these genome sequences were determined by GLIMMER (40,201) which showed 2118 and 2175 protein coding genes in strains NEM316 and 2603V/R respectively. ORF's can be placed into families and super-families on the basis of similarity with known protein sequences using Markov models such as those used in the programs PFAM (11) and TIGRFAMS (78). This allows characterisation of genetic features such as the levels of surface genes and potential virulence genes (68,232).

Comparison of multiple genomes is more informative. The genome sequences of NEM316 was compared to the previously sequenced *Streptococcus pneumoniae* strain TIGR 4 (233) and *Streptococcus pyogenes* strain M1 (60) and showed that out of the 2118 protein coding genes 1173 and 1139 were shared in *S. pneumoniae* TIGR4 and *S. pyogenes* M1 respectively. Strain 2603V/R shared 1236 genes with *S. pneumoniae* and 1285 genes with *S. pyogenes*. Comparison to other genomes can also reveal unique genes with links to pathogenesis. For example, analysis of strain 2603VR revealed that genes encoding the CPS capsule and genes essential for *S. agalactiae* β-haemolytic activity are unique to *S. agalactiae*. Analysis of multiple genomes from the same species allows the study of the genetic basis of pathogenicity by identifying unique and core genes. Core genes between strains of the same species reveal the backbone of the species and pathogenic elements common to the species (for example, the CPS capsule), while unique genes give clues to the variation in pathogenicity between strains. The core genome of *S.agalactiae* has been defined by a number of methods to be discussed later. Tettelin et al. placed the core genome between 1,750–1,841 (~85%) genes (231) and Lefebure et al. placed the number of core genes at 1686 (~80%). Konstantinidis et al. showed that strains 2603V/R and NEM316 share 87.5% of their genes (120) in his work using Average Nucleotide Identity (ANI) to develop a genomic method of taxonomic classification. ANI is a measure of species diversity, it is the average identity shared between genes in the core genome and therefore is a measure of how rapidly genomes are evolving and it has been shown that the phylogeny of genes with nucleotide identity values that correlate to the ANI accurately reflect whole genome phylogeny (119). Additional information from the genome of an organism can be discovered by performing genome alignments using tools such as the Artemis Comparison Tool (ACT) (28) to show chromosomal rearrangements and/or novel genomic islands. For example, Tumapa et al. studied genome plasticity in *Burkholderia pseudomallei* and identified 5 genomic islands (235).

1.2.3 Virulence Factors

There are already a large number of known virulence factors in GBS which can be grouped into several classes with differing roles in pathogenicity of the organism. The largest group of virulence factors are adhesins which are cell surface components involved in binding to host cells (37). GBS adhesins include fibrinogen binding proteins A and B (fbsA and fbsB) (212). Both bind human fibrinogen (208) but fsbA is attached to the GBS cell wall whereas fbsB is secreted (75,96). The pilus islands I and II are genomic islands containing genes which encode and regulate pilus structures (127,195,229) that have been shown to adhere to endothelial cells of the blood brain barrier (148) and the Lmb lipoprotein which has been shown to aid invasion of damaged epithelium (220). GBS also contains a number of exoenzymes which are secreted enzymes that break down extra-cellular products including hyaluronidase which degrades hyaluronic acid present in the ground substance of connective tissue (93) and streptococcal enolase, a glycolytic enzyme which aids GBS binding to plasminogen (173).

A number of GBS virulence factors are categorised by their induction of an immune response. The immunoreactive antigens are the C, X and R protein antigens used in serotyping (117) and have diverse functions (24,193,224,248). They include the α-C protein which mediates internalisation of GBS in human cervical cells (19), the α-like protein which is highly similar to the α-C protein except it carries an IgA-binding region similar to the one found in β-C protein (122), the β-C protein which may aid in preventing opsonophagocytosis (6,98), resistance to proteases (rib) protein which is carried by most serotype III GBS isolates and is involved in immune evasion (224,248) and the surface immunogenic protein (sip) which is of unknown function but is highly conserved in GBS (24,193).

GBS has also been shown to contain genes that are involved in immune evasion. For example the capsular polysaccharide locus prevents deposition of complement factor C3b and shields immunogenic proteins on the cell surface through molecular mimicry of mammalian sugar epitopes (4,46). Additionally C3-degrading protease and C5a peptidase inactivate C3 and C5a complement factors respectively (5,35) and serine protease is a surface expressed protein

involved in folding and maturation of secreted proteins and inhibition of genetic competence as part of the CiaRH two component regulatory system (94). Finally, haemolysin is a potent cytotoxin affecting a broad range of host cells (46) and is responsible for β-haemolysis of red blood cells when GBS is plated on blood agar (56).

1.2.4 Regulation and Nucleotide Repeats

GBS can exist in multiple locations within the body and expresses different proteins at different stages in its pathogenesis therefore, different genomic expression profiles can be seen when the bacteria is subjected to different conditions. Mereghetti et al. showed that incubation of GBS isolates in human blood rapidly changed the transcription profile of the organism when compared to growth in laboratory rich medium. Proteins involved in interaction with the host coagulation/fibrinolysis system and proteins involved in bacterial-host interactions were rapidly up-regulated. Additionally, extensive transcript changes also occurred for genes involved in carbohydrate metabolism, probably indicating the relative scarcity of carbohydrates in human blood compared to laboratory media, including multi-functional proteins and regulators putatively involved in pathogenesis (161). Changes in the transcriptome of GBS can even be seen in different growth stages under laboratory conditions with an up-regulation of genes involved in virulence factor production and utilization of alternate carbon sources (214). The response of GBS to growth in human amniotic fluid has also been studied and showed that the majority of up-regulated genes were involved in amino acid and nucleotide production and metabolism, again reflecting the levels of available nutrients in laboratory medium compared to practically any source where the bacteria will be found in nature. Additionally, multiple virulence genes such as adhesins, capsule genes, hemolysin and IL-8 proteinase were shown to be up-regulated potentially affecting host-pathogen interactions (213).

Having shown that GBS responds to its environment as would be expected, it is worth considering the mechanisms in place to ensure this regulation. Currently, only a limited number of regulatory

methods are understood (187) but a number of genes important in virulence can be shown to be regulated by a variety of processes. Samen et al. showed using gene knock out models that the *rovS* gene is involved in regulating genes involved in attachment to human epithelium as well as a number of other key virulence genes (202). Rozhdestvenskaya et al. showed that the BgrR/S two component regulatory system led to decreased β-antigen expression and also led to reduced virulence properties in *S. agalactiae* (197). Rajagopal et al. have shown that Stk1 positively regulates transcription of a cytotoxin, β-haemolysin/cytolysin that is critical for survival of GBS in the bloodstream and for resistance to oxidative stress (189). To summarise, GBS seems to have a repertoire of regulatory mechanism that deal with changes in environmental conditions, which is not unsurprising considering the number of different environments GBS can be isolated from.

However, response to the human immune system can often involve other methods of genetic regulation. For example Short Sequence Repeats (SSRs) are simple DNA sequence repeats made up from motifs of 1-6bp that are repeated up to a dozen times (246) and there is evidence that these regions play a role in regulation of gene expression (74). Mononucleotide repeats (MNRs) have been linked to virulence genes due to their role in phase variation by slipped strand mispairing. Evidence for this includes the fact that SSR tracts above 3bp are heavily over represented in coding regions of the genome of *E. coli* (74). Also, on the basis of analysis of 81 bacterial and 18 archaeal genomes, Orsi et al. suggests that homopolymeric repeats represent a general regulatory mechanism in prokaryotes since homopolymeric repeats are over represented in the first 10% of genes. This suggests a role in genomic regulation by slipped strand mispairing since it would be beneficial to the organism that transcription is terminated earlier in the gene to avoid expending resources on a non-functional transcript (171).

1.3 Tools and Technology

1.3.1 Bioinformatics

1.3.1.1 Phylogenetics

Phylogenetics is defined as the study of evolutionary relatedness among groups of organisms which is discovered through molecular sequencing data and morphological data matrices. The term *phylogenetics* is of Greek origin from the terms *phyle/phylon* meaning "tribe/race" and *genetikos* meaning "relative to birth". Methods for estimating phylogenies include distance-based methods such as neighbour-joining and Unweighted Pair Group Method with Arithmetic Mean (UPGMA), maximum parsimony, Bayesian phylogenetic inference and Maximum Likelihood.

The distance based methods rely on clustering genetic distance values and therefore require either a multiple sequence alignment or a distance matrix as an input (164). Parsimony is a method that uses character classes and therefore can be used with both sequence and morphological data. It operates by evaluating candidate phylogenetic trees according to an explicit optimality criterion which provides a measure of the fit of the data to a given hypothesis, the tree with the most favourable score is taken as the best estimate of the phylogenetic relationships of the included taxa. DNA can easily be divided into character classes since at any given position in an alignment the character can be any IUPAC base (218).

Alternatly, eBURST (59) can be used to determine relationships between molecular typing data that defines isolates as strings of integers such as MLST allelic profiles. It works by attempting split allele types from sequence typing schemes into groups and then identify the founding sequence type of each group. The algorithm then predicts the descent from the predicted founding genotype to the other genotypes in the group displaying the output as a radial diagram, centred on the predicted founding genotype. Since it used allele type data eBurst is particularly good at identifying recombination events. But since it does not use sequence data directly it assumes that all allele types are equally related which is a weakness for this particular method (59).

Maximum Likelihood generates the most likely tree given the confines of a previously selected model of evolution (258). Therefore selection of the model of evolution is crucial to creating good maximum likelihood trees. The ModelTest software (182) can select the best possible evolutionary model using the Akaike Information Criterion. The maximum likelihood process itself generates all possible trees and assesses the log likelihood of each tree against the sequence data given the selected model of evolution. The most statistically significant log likelihood value is considered the most likely tree. Bayesian phylogenetic approaches are computationally similar to maximum likelihood approaches with the main difference being the addition of prior and posterior probabilities. These are assumptions about the data made before generating a tree. Additionally, maximum likelihood generates one optimised tree whereas Bayesian methods generate multiple trees and each tree generated is used to generate a final consensus tree. That is, rather than the maximum likelihood "hill climbing" algorithm to generate the best possible tree, the Bayesian method uses a Markov Chain Monte Carlo (MCMC) algorithm of a large sample of highly probably trees to generate a consensus tree (Oliver Gascuel and Manolo Gouy, Personal Correspondence April 2009). It is also worth considering the "curse of dimensionality" in regards to this data. Since the number of sequences to consider is potentially rather large. The "curse of dimensionality" refers to issues that present themselves when analyzing data in high-dimensional spaces (such as data with a large number of variables). Essentially, high dimensional data may not form sufficiently defined clusters making data difficult to resolve. This becomes more of an issue as sample size increases and may be an issue with the number of sequences being analysed here. However, there are steps that can be taken to reduce the likelihood that high dimensionality data will cause problems.

Firstly we can consider the data we are inputting to the model itself. Since even low dimensional data can be turned into high dimensional data by the addition of duplicates only sequences that are representative of sequence types were used here therefore limiting the dimensionality.

Also, the curse of dimensionality is essentially a problem of clustering too large of a number of samples in the same space and subsequent failure to resolve these clusters the selection of more

variable sequences should somewhat overcome the issue by creating a larger space for variables to exist in.

Secondly, the selection of the Maximum Likelihood model somewhat overcomes this issue since the MCMC random walk based sampling method moves from point to point which limits the issues caused by large uniform clusters. This does make high Dimensionality data more computationally intensive as more steps are required before finding the optimal likelihood but since 1) computational power is in abundance and 2) steps have already been taken to reduce the dimensionality of this data this is not expected that high dimensionality data will cause significant problems. Additionally, since all analyses were confirmed by bootstrapping, analysis problems caused by high dimensionality data could be identified by poor bootstrap support for points on the tree.

1.3.1.2 Identification of the Core/Variable Genome

The core genome is defined as the pool of genes shared by all strains of the same bacterial species (159) but the term has also been used to describe genes shared within a genus (132).The principle behind establishing the core genome is relatively simple although in practice it may be more difficult. DNA or protein sequences within a genome are compared to gene or proteins sequences from one or more other genomes. Comparison is based on levels of homology which are calculated using either alignment algorithms (e.g. FASTA or ClustalW) or the BLAST algorithm. Despite the underlying mechanisms being similar different programs use different parameters to decide what constitutes a good alignment, each set of parameters are known as scoring matrices. Tettelin et al (231). has used three methods to establish the GBS core genome, firstly Smith and Waterman protein searches on all of the predicted proteins by using the SSEARCH program (which uses alignments created using Smith-Waterman's scoring matrix). Secondly a DNA search of all of the predicted ORF's of a strain against the complete DNA sequence of the other strain using the FASTA program and finally a translated protein search of all of the predicted proteins of a strain

against the complete DNA sequence of the other strain using the TFASTY program. Three separate methods were used to ensure reproducibility of bioinformatic results. When any of the three methods identified an alignment with a minimum of 50% sequence homology, over 50% of the protein/gene length, that gene was considered core (231). However using three different methods and considering genes picked by any one method as core, rather than taking genes picked by all three methods, may lead to over overrepresentation of the core genome.

Konstantinidis used a reciprocal best match BLAST approach to establish the core genome (119). In this method every gene in a reference genome is BLASTed against one or more query genomes using homology and length criteria. Genes that meet these criteria in the query genome are BLASTed back against the reference genome. If the set criteria are met in both searches then the gene is considered core and is used in further analysis. This method is considered robust and reproducible and it has been used previously to establish the core genome of four bacterial groups *Escherichia coli*, *Salmonella* spp., *Shewanella* spp., and *Burkholderia* spp (119).

The variable genome is defined as everything in the genome which is not considered core and can easily be identified after a core genome analysis has been performed. Studying the variable genome allows identification of genes that differ between genomes and may identify potential virulence factors present in pathogenic strains that are not present in less/non pathogenic bacteria.

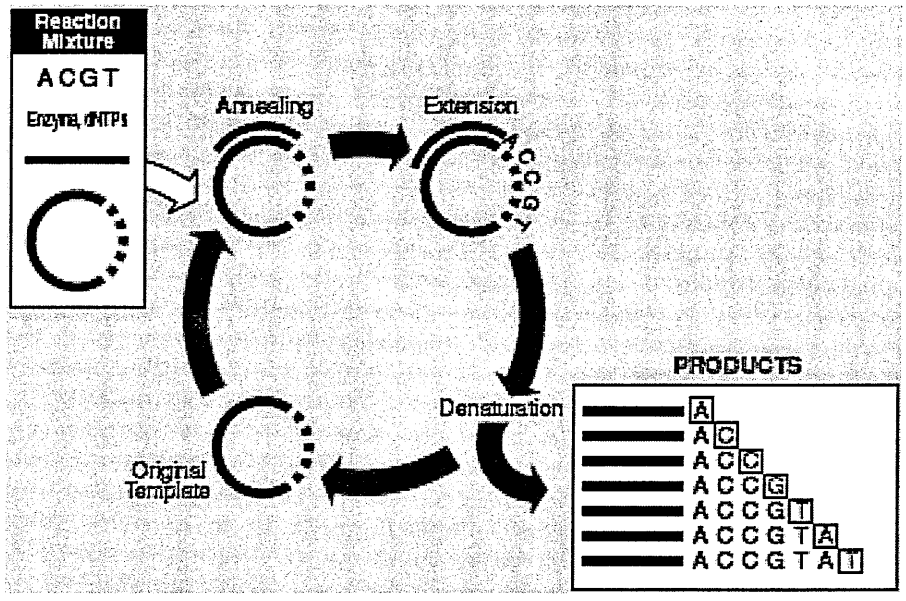
Recent advances and reduction in the cost of whole genome sequencing technologies has made sequencing of bacterial genomes quicker and cheaper and therefore has led to far higher volumes of genome sequences being available making comparative genomic analysis far more important and easier to perform. Because of this abundance of data new tools for analysis of the core genome are being developed, for example Panseq is an online tool to identify the core and accessory genomes using the BLASTn program of any given set of genomes (124) and CoreAligner which relies on creating alignments of orthologous groups to identify common regions within genomes (237). As whole genome bacterial sequencing becomes more common in research it is

highly likely that software to perform comparative genomic analysis between next generation sequenced genomes will become standardised and even supplied as part of the standard software supplied with next generation sequencing platforms.

1.3.2 Sequencing

1.3.2.1 Sanger Di-deoxy Sequencing

Figure 1.9: Representation of the Sanger sequencing process



Source: Applied Biosystems

Di-deoxy sequencing is the most established sequencing method. First proposed by Sanger et al. in 1975 it works via the chain termination method (203). DNA amplified by PCR is placed into a reaction mix containing DNA polymerase, an oligonucleotide primer, a mix of deoxy and di-deoxy nucleotides and is subjected to cycle sequencing. As in PCR, DNA polymerase will extend a chain of nucleotides complementary to the template DNA. When a di-deoxy nucleotide is incorporated into the sequence further extension cannot occur. Within the reaction, this will occur at each nucleotide creating fragments at different lengths based on where a di-deoxy nucleotide is

incorporated and will occur a sufficient number of times for visualisation of each fragment on a gel. Traditionally, one di-deoxy nucleotide is used per reaction and each reaction (A, G, C and T) is ran on one lane of a gel allowing the sequence to be read from the four lanes on the basis of size of DNA fragments (203,204). Sequencing has now been automated and each di-deoxy nucleotide is tagged with a different fluorescent dye and all 4 termination reactions are performed in the same reaction, analysed by capillary electrophoresis using an automated sequencer and results in a chromatograph of peaks corresponding to each base.

Despite di-deoxy sequencing being the gold standard for sequencing and chromatographs being accepted by all major sequence repositories, there are disadvantages to the method. The throughput is lower than next generation sequencing technologies. Sequencing whole genomes using sanger sequencing would require dedicated centres with up to 100 sequencing machines running 24 hours a day 365 days a year using a highly automated template preparation process, without this set-up, sequencing even the smallest genome would be prohibitively expensive and require massive manpower (81). It is also not the best method for sequencing short DNA fragments.

1.3.2.2 Sequencing by Synthesis (SbS)/Pyrosequencing

Sequencing by synthesis is a method for rapidly sequencing short sequences of DNA of up to 100bp (157) although this length has been improved using whole genome sequencing platforms. This method uses either 3 or 4 enzymatic steps depending on the platform. The four nucleotides (dATP, dTTP, dCTP and dGTP) are added sequentially for each base being sequenced. When a nucleotide is incorporated into the extending DNA it releases inorganic pyrophosphate (PPi) which is converted to adenosine triphosphate (ATP) by ATP sulfurylase. The released ATP causes luciferase to emit light which can be detected. The level of light emitted indicates how many nucleotides are incorporated (i.e. when dATP is introduced, if there are 5 T nucleotides in the complementary sequence then 5 dATP nucleotides will be incorporated emitting 5 times as much

light as the incorporation of one). The fourth enzyme is apyrase which is used to degrade unincorporated dNTPs and ATP in the Biotage system, other pyrosequencing based systems such as the Roche 454 platform replace this fourth enzyme with a washing step (157). Generally in pyrosequencing, sequencing occurs at a speed of ~1bp per minute (105).

Pyrosequencing has been used to type bacteria. Luna et al (143). used pyrosequencing to identify atypical bacterial isolates within a children's hospital by sequencing the highly variable V1 and V3 regions of 16S rRNA gene and Nygren at al (169). used pyrosequencing to differentiate *Bordetella* using a section of the Pertussis Toxin promoter region to differentiate virulent *Bordetella pertussis* from less virulent *B. parapertussis* and *B. bronchiseptica*. Sequencing of short fragments was shown to identify mutations in *B. parapertussis* and *B. bronchiseptica* that rendered the promoter region non-functional and therefore prevented expression of Pertussis Toxin. Pyrosequencing was selected since a region of the Pertussis Toxin promoter showed sequence compression when sequenced by Sanger sequencing due to strong secondary structure formation. Finally, Wroblewski et al (256). developed a rapid method to characterise *Clostridium difficile* strains using pyrosequencing to sequence fragments of the genes encoding toxins A (*tcdA*) and B (*tcdB*) and the binary toxin genes (*cdtA* and *cdtB*), and detect common deletions in the *tcdC* gene which is thought to be involved in negative regulation of toxin gene expression.

1.3.2.3 Second Generation Whole Genome Sequencing

Advances in whole genome sequencing technology have allowed sequencing of a large number of bacterial genomes leading to increased understanding of bacterial diversity. Currently, the sequencing technologies used are referred to as second generation platforms and include the 454 Genome Sequencer FLX (Roche Applied Science), the Illumina (Solexa) Genome Analyzer and the ABI SOLiD System (Applied Biosystems).

The 454 and Illumina platforms both use sequencing by synthesis chemistry in parallel reactions to generate large numbers of short sequencing reads (each read for the 454 platform is ~400bp

whereas the illumina is ~150bp). The 454 platform involves library creation by fragmenting genomic DNA to approximately 300-800bp. Short adaptors are added to the 5' and 3' ends of each fragment. These adaptors are used to attach DNA fragments to propriety DNA capture beads. Each bead will have one fragment attached to it and each bead acts as an individual micro-reactor for emulsion PCR meaning each bead contains DNA amplified from the individual fragment attached to it. Each amplified bead is then added to PicoTiterPlate device for sequencing. The size of the wells means each well will only contain one bead and in each well an individual sequencing by synthesis reaction is performed allowing sequencing of 1,000,000 individual reads with the GS FLX System and 100,000 individual reads with the GS Junior System per 10-hour instrument run. The Illumina platform uses the same steps of library creation and clonal amplification of DNA fragments but sequencing and amplification of these fragments is performed inside flow cells using standard sequencing by synthesis reagents.

The ABI SOLiD platform uses sequencing by ligation chemistry where DNA fragments are ligated to beads and clonally amplified. The beads are then enriched to separate the beads with extended templates from undesired beads. The template on the selected beads undergoes a 3' modification to allow covalent attachment to a glass slide. Sequencing then occurs by ligating a 5bp probe with a florescent tag complementary to the first 2bp of sequence to the template strand. Following measurement of the florescent tag to determine the first 2bp of the 5bp probe the florescent tag is cleaved and the next 5bp tag is bound to the next 5bp of the template sequence. The ligation reaction is repeated as required. Following the ligation reactions the extension product is removed and the sequencing primer is replaced by a primer complementary to N-1 of the first primer for a second round of ligation. This allows sequencing of the 2/5 bp of the probe which is out of frame by 1bp to the first set of ligation reactions. This process of resetting the primer to the n-1 position of the previous primer and performing a new series of ligation reactions is repeated 5 times to sequence all bases of the template DNA.

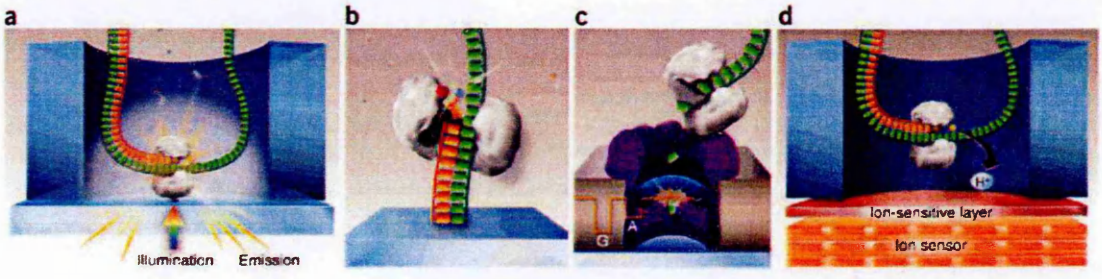
Having the capability to sequence such a large amount of DNA has allowed researchers to perform experiments that would have been prohibitively expensive even a few years ago. For

example, Morelli et al. used 454 sequencing technology to sequence seventeen *Yersinia pestis* genomes to identify 933 SNPs between isolates that where then screened using Sequenom MassARRAY SNP typing in 286 isolates (163). Holt et al. used Illumina and 454 technology to sequence the genomes of 19 *Salmonella enterica* serovar Typhi isolates to show that evolution in the Typhi population seems to be characterized by ongoing loss of gene function and discovered a lack of evidence for antigenic variation driven by immune selection in contrast to strong adaptive selection for mutations conferring antibiotic resistance (90). Finally, Gulig et al. used the SOLiD system to sequence four strains of *Vibrio vulnificus*, two strains came from a virulent population and two strains were environmental strains. This revealed eighty genes specific to the virulent strains (73).

1.3.2.4 Third Generation Whole Genome Sequencing

There are currently four third generation sequencing platforms that use single strand DNA sequencing the Pacific Biosciences Single Molecule Real Time (SMRT) (figure 1.10-a), the Life Technologies FRET sequencing platform (figure 1.10-b), the Oxford nanopore sequencing platform (figure 1.10-c) and the Ion Torrent sequencing platform (figure 1.10-d).

Figure 1.10: Four third generation sequencing technologies



Source: Nature Biotechnology

The Pacific Biosciences Single Molecule Real Time (SMRT) achieves single-molecule sequencing by detecting nucleotides which are fluorescently labelled at the terminal phosphate when they are held in place by a polymerase during incorporation.

The Life Technologies FRET sequencing platform uses fluorescently labelled nucleotides, a DNA polymerase modified with a quantum dot and DNA template molecules immobilized onto a solid surface. During incorporation events, energy is transferred from the quantum dot to an acceptor fluorescent moiety on each labelled base.

The Oxford Nanopore's platform uses an exonuclease coupled to a modified α-hemolysin nanopore which is positioned within a lipid bilayer. As DNA is directed through the α-hemolysin each nucleotide is cleaved sequentially. As single bases fall into the pore, they transiently interact with the cyclodextrin moiety which disturbs current through the nanopore in a manner which is unique for each base. (198).

The Ion Torrent sequencing platform uses a semiconductor-based high-density array of microwell reaction chambers positioned above an ion-sensitive layer and an ion sensor. As single nucleotides are sequentially incorporated the release of hydrogen ions is measured to identify the incorporated nucleotide.

1.3.2.5 Sequence Analysis by MALDI-TOF MS

Matrix Assisted Laser Deabsorption/Ionisation Time of Flight (MALDI-TOF) mass spectrometry has been used for sequencing and Single Nucleotide Polymorphism (SNP) detection. MALDI-TOF is a form of mass spectrometry which uses laser pulses directed at a crystallised mix of analyte and a chemical matrix molecule. When a laser pulse is directed at this mix, energy is transferred to the matrix molecule causing the analyte and matrix to be ionised. The ionised analyte is carried by an electromagnetic field to a detector. The time it takes the ionised analyte to reach the detector is used to calculate its mass (87). There are two Sequenom applications used in genotyping DNA, the

iSEQ system which sequences loci using base specific cleavage and the iPLEX system which is used for SNP genotyping.

The iSEQ system uses RNA fragments generated through base specific cleavage of reverse transcribed PCR amplified DNA (225). Cleavage occurs at the C and T (U) bases in the forward and reverse strands of RNA giving 4 reactions per sequence. The size of fragments in each reaction is then compared to a reference sequence. Variations in sequence are discovered by identifying which fragments in the reference sequence do not appear as expected, the software then calculates the new sequence on the basis of mass fragments that are not matched to the reference sequence. Since there are 4 cleavage reactions per sequence all nucleotide variations are proved since they are replicated within the experiment. The iSEQ system has been used to perform MLST typing of *N. meningitidis* and produced results consistent with di-deoxy sequencing (91) and to characterise SNPs in the genome of JC polyomavirus (12).

The iPLEX system is used for genotyping SNPs and uses a primer designed immediately upstream of a SNP. The primer is then extended by one nucleotide complementary to the template strand and differences in mass between the extended primers allow SNP typing. Recently, this application has become more popular than the iSEQ since it can genotype up to 40 SNPs per well in a 384well plate, allowing up to 15360 SNPs genotyped per plate making it ideal for studies of SNPs located through whole genome sequencing. For example, Morelli et al. (163) used 454 sequencing to sequence 17 *Yersina pestis* genomes and using the data generated identified 933 SNPs which were then profiled in 286 *Y. pestis* using the iPLEX platform. Clawson et al. sequenced pooled DNA from 91 human *E. coli* O157.H7 infections and 102 *E. coli* 0157.H7 isolates from cattle which identified 16,218 putative polymorphisms. From those, 178 were genotyped using the iPLEX platform which allowed differentiation of human and bovine isolates (34).

1.4 Aims and Objectives

GBS normally exists as a commensal organism of the gastrointestinal and genitourinary tracts but in a small number of cases the bacteria can be passed to the neonate during childbirth to cause systemic disease. To aid in the understanding of why some bacteria cause infection and others do not, a method of generating DNA profiles which accurately reflect phylogeny and give indications of potential virulence are required. The aim of this project is to develop a two component molecular profiling system to address these issues. The first component will look at sections of the genome that are evolving at the same rate of the organism as a whole and the second component will look at regions of the genome containing MNR repeat regions around or within virulence genes.

The first component of this profiling system will be developed by:

- Adapting a method which was developed to select targets for sequence typing at the genus level (119) to a select targets at the strain level.
- This method will use reciprocal BLAST to identify the core genome which will be analysed to calculate the Average Nucleotide Identity (ANI) of the core genome which will be correlated to the evolutionary distances generated from Maximum Likelihood analysis of each core gene
- The assumption is that genes with a strong correlation to the rate the genome is evolving at the average rate will more accurately reflect the evolution of the organism.
- Selected targets will be sequenced from a collection of UK clinical isolates and compared to sequence type and levels of diversity of the loci of the GBS MLST scheme (103).

The second component of this project will look at the presence/abundance of Mononucleotide Repeats (MNRs) in the genome as:

- MNRs have been shown to be a genomic regulator in both coding (74) and non-coding (184) DNA meaning analysis of MNR repeats within or around virulence associated genes may allow predictions of virulence.

- Previous studies have demonstrated that these MNR containing regions could be suitable profiling markers (42).
- Within the core-genome MNRs will be assessed by developing Perl workflows to assess the presence, abundance and homology between MNR repeats, focusing specifically on virulence genes.
- MNR repeats will also be assessed within non-coding DNA, focusing on non-coding DNA at the 5' and 3' of known virulence factors.
- Selected targets will be assessed as profiling markers and results compared to the existing MLST scheme and first component of the profiling system

In the final stage, analysis will be carried out to find the optimal way to combine the two components to develop an optimal profiling system for GBS which is able to accurately reflect the evolution of the organism, discriminate more accurately than the currently used MLST scheme and provide indicators of the virulence of the organism.

Chapter 2

Materials and Methods

2.0 Materials and Methods

2.1 Laboratory Methods

2.1.1 Strain Collection

This study uses a collection of 134 GBS isolates consisting of 48 reference strains (including 8 genome sequenced isolates) and 86 clinical isolates from the UK. The collection represents capsular serotypes IA, IB and II-VIII and examples of the immunoreactive antigens C, R and X as shown in Appendix 9.1. A summary of the serotypes of the isolates is shown in table 2.1.

Table 2.1: Serotypes of clinical isolates used in this study

Serotype	Frequency
IA	18
IB	14
I (Other)	3
II	17
III	18
IV	12
V	14
VI	5
VII	7
VIII	4
NT	22
Subtypes	
C	19
R	3
X	3

2.1.2 Bacterial Culture

Each strain was plated from Protect bacterial preservation system beads (Technical Science Consultants Ltd., Heywood, UK) stored at -80°C onto Columbia agar with Defibrinated Horse Blood (HPA Media Department, London, UK), Plates were first checked for purity and then streaked to provide single colonies and grown overnight at 37°C. Single colonies were then picked and sub-cultured overnight for DNA extraction.

2.1.3 DNA Extraction

DNA extraction was performed with the Qiagen DNeasy Blood & Tissue kit using a modified Gram positive extraction protocol (Qiagen Ltd, Crawley, UK). A 10µl loop of cells were taken from the agar plate and re-suspended in 1ml phosphate buffered saline (PBS) (HPA media department) and centrifuged for 10 minutes at 7500rpm using a Sigma 1-15p centrifuge (Scientific Laboratory Supplies Ltd., Nottingham, UK). The supernatant was removed and the pellet was re-suspended in 180µl enzymatic lysis buffer consisting of 20mM Tris·Cl (pH 8.0), 2 mM sodium EDTA, 1.2% Triton® X-100, lysozyme at 20mg/ml and mutanolysin at 50µg/ml which was incubated at 37°C for 30 minutes. Twenty five microlitres proteinase K and 200µl Buffer AL were then added, mixed by vortexing and incubated for 1 hour at 56°C. Following incubation 200µl 96-100% ethanol (Sigma-Aldrich Co.Ltd, Gillingham, UK) was added and mixed by vortexing. The mixture from this step was added to a DNeasy Mini spin column placed in a 2 ml collection tube and centrifuged at 6000xg (8000 rpm) for 1 minute after which the flow-through and collection tube were discarded. The spin column was then placed into a new collection tube and 500µl of buffer AW1 was added to the spin column which was centrifuged at 6000xg (8000rpm) for 1 minute. Again, the collection tube was discarded and the spin column was placed in a new collection tube and 500µl of buffer AW2 was added to the spin column which was centrifuged for 3 min at 20,000xg (14,000 rpm). The collection tube was again discarded and the spin column was placed into a 1.5ml microcentrifuge tube (Eppendorf UK Ltd, Cambridge, UK) with the cap removed. 100µl buffer AE was added directly to the spin column membrane and centrifuged at 6000xg (8000rpm) for 1 minute, this stage was then repeated to ensure maximum DNA elution. The 200µl eluted product was then transferred to a screw-topped microcentrifuge tube for storage at -20°C.

2.1.4 DNA Quantification

Extracted DNA was quantified using the NanoDrop ND-8000 spectrophotometer (Labtech International Ltd, Ringmer, UK). Three microlitres of DNA from each sample was transferred to a 96 well skirted PCR plate (Abgene, Epsom, UK). Two microlitres of water was used to initialise the instrument, this was cleaned off using lint free wipes (Kimberly-Clarke Professional) and replaced with 2µl of Qiagen DNeasy buffer AE and the instrument was "blanked" i.e. a sample of elution buffer was used to provide a set of spectral measurements that are assumed to contain no DNA. DNA samples were then added to the 8 channels and the absorbance at frequencies A230, A260 and A280 were measured. The A260 measurement is used by the software to calculate the concentration of DNA (measured in ng/µl) by using the Beer-Lambert equation which states that A = E x b x c where A is the absorbance, E is the extinction coefficient, b is the length of the wave path and c is the analyte concentration. The A260/280 ratio is expected to be around 1.8 for pure DNA, if the ratio is lower, it could indicate the presence of protein, phenol or other contaminants that absorb strongly at or near 280 nm and the A260/230 is a secondary measure of DNA purity and is typically between 1.8-2.2, lower values may indicate the presence of co-purified contaminants. After measurements are taken the 8 channels are cleaned again and this stage is repeated until all samples are processed.

2.1.5 PCR

All PCR reactions were performed using Qiagen HotStar Taq 2x Master Mix (Qiagen) supplemented with 1.0mM $MgCl_2$ (final $MgCl_2$ concentration of 2.5mM). Primer concentrations were 0.2µM for all primer pairs except the T7/SP6 tagged *glnA* sequencing primers used in the Sequenom amplification reaction which were used at a concentration of 0.3µM. Early optimisation of the MLST primer pairs showed that cycling conditions of an initial 15 minute denaturation followed by 30 cycles of a 30 second denaturation at 95°C, a 30 second annealing step at 55°C and a 1 minute elongation step at 72°C was optimal. After 30 cycles there was a 7 minute final elongation stage at 72°C. These conditions were used as standard for all further primer pairs unless they did not produce sufficient product. The non-coding primer pairs (see table 2.2) did not produce sufficient product until the annealing temperature was raised to 60°C.

2.1.5.1 Primers

All primers were synthesised by Eurofins MWG Operon (Eurofins MWG Operon, Ebersberg, Germany). Primers are dispatched as lyophilised powder and are made up to a stock concentration of 100pmol/µl and stored at -20°C. For use, PCR primers are made to a working concentration of 0.2 or 0.3µM (see above). The sequence of each amplification primer used is shown in table 2.2.

Table 2.2: PCR amplification primers

Primer	Scheme	Forward (5' - 3')	Reverse (5' - 3')	Product Length (bp)
adhP Amplification	MLST	GTTGGTCATGGTGAAGCACT	ACTGTACCTCCAGCACGAAC	672
atr Amplification	MLST	CGATTCTCTCAGCTTTGTTA	AAGAAATCTCTTGTGCGGAT	627
glcK Amplification	MLST	CTCGGAGGAACGACCATTAA	CTTGTAACAGTATCACCGTT	607
glnA Amplification	MLST	CCGGCTACAGATGAACAATT	CTGATAATTGCCATTCCACG	589
pheS Amplification	MLST	GATTAAGGAGTAGTGGCACG	TTGAGATCGCCCATTGAAAT	723
sdhA Amplification	MLST	AGAGCAAGCTAATAGCCAAC	ATATCAGCAGCAACAAGTGC	646
tkt Amplification	MLST	CCAGGCTTTGATTTAGTTGA	AATAGCTTGTTGGCTTGAAA	859
valS	Coding	ACCAAAATACAATCCTGCCG	AAGCACCCTCAACATCCTTG	581
mmuM	Coding	TCGCTTGACTGTTCAGTTGG	AACAACCTCCGACAACTTGG	629
obgE	Coding	CTAAAGGTATGCACGGTCGAG	TTTTCTTTGAATGCTGCCAA	676
SAG1894	Coding	TGCCAAGACTACTGTTTGGAGA	TCTAGTGCGAAACGATTTTTGA	605
SAG0025	Coding	CTACTACATTCCAGGTCAAGC	GAAGATGATATTTGCCTTAGC	537
SAG0027	Coding	TGTGTTGCCATGTGTGTCAA	ATTTCAAACATTTCTTCGTGC	590
SAG0043	Coding	TAAAGCCTACATCGAAGAGCA	CTTTCTCGTAATCGAGCGGAT	601
SAG0047	Coding	GTACCCACGGTGTTCACGC	CCAAAAGTTGATTCCATATTGCGC	590
cpsL	Coding	GCACCATTAGTTGGTTTTTCT	TCTCCTCCCATTATTTGAGC	626
cylB	Coding	AGGTGCCTTTGGAGTTATGG	TAGAGACAGTGGCTTCGTTGG	500
SAG0032 Non-Coding	Non-Coding	TTGGAATGCAATGCCAGAT	GGCATAACACCTCCACTTTCA	435
SAG0043 Non-Coding	Non-Coding	GCAGGGCAGGACAAAATCTA	TGGGCAGAGACGACTTTCTT	520
SAG0106 Non-Coding	Non-Coding	ACGAAGAAGACGATGAAGAGGA	ACCTTTAGCCCACGGTTTTT	420
SAG0649 Non-Coding	Non-Coding	CCTAAAACAGGGGGAATTGG	CCGTTGCAGATTGTCCTCTT	469
SAG1768 Non-Coding	Non-Coding	TGTGGTTTGACTGACATGCTG	TCAAACTGTTGACGGTAACCAA	402
SAG2143 Non-Coding	Non-Coding	ACCTTTAGCAGAGCCACCTAT	TCACCAAAATGATGGGACTT	431
SAK_1320 Non-Coding	Non-Coding	AACTGACATGCCTTGCGTAG	TGACCACTTTCTTCTTGGGATT	411

2.1.5.2 PCR Set-Up Using the Corbett Robotics CAS4200

The CAS4200 (Qiagen) is an automated high throughput PCR laboratory robotic system which was used to prepare PCR reactions for Sequenom iSEQ sequencing and Sanger Di-Deoxy sequencing. The program Robotics4 (Qiagen) was used to program the plate layout to configure the master mix ingredients, sample locations (and in some cases the sample concentrations) and determine the order of reagent, reaction mix and/or sample addition.

Additionally the CAS4200 was used to normalise DNA concentrations to ensure a constant 5ng/µl DNA concentration in Sequenom PCR reactions.

2.1.6 Gel Electrophoresis

Gel electrophoresis was performed in a Sub-Cell Model 96 Cell Submerged Horizontal Electrophoresis Tank using a 25x15cm gel caster and two standard 50 well combs (Bio-Rad Laboratories Ltd., Hemel Hampstead, UK). Each gel was made to a thickness of 5mm using 2% UltraPure agrose (Life Technologies, Paisley, UK) in 0.5x TBE (Life Technologies). A BioMarker ext plus 50-2500bp ladder (Cambio Ltd., Cambridge, UK) was used to size DNA products. One microlitre of PCR reaction product was added to 5µl 1x BlueJuice loading buffer (Life Technologies) before being run for 1 hour at 120V with a maximum current of 100mA in 0.5x TBE running buffer (Life Technologies). Gels were stained in 500ml 0.5mg/l ethidium bromide solution and visualised using a Gel-Doc 2000 (Bio-Rad).

2.1.7 Sequenom iSEQ Re-sequencing

2.1.7.1 Experimental Design

2.1.7.1.1 PCR Primers and Tags

Due to the reverse transcription stage in the iSEQ workflow PCR primers need to be tagged with reverse transcriptase promoter sequences. The primers used were the MLST sequencing primers from the pubMLST website (http://pubmlst.org/sagalactiae/info/primers.shtml) where the forward primers are tagged with a T7 polymerase promoter and the reverse primers tagged with a SP6 polymerase promoter (see table 2.3). All primers where synthesised by Eurofins MWG Operon and prepared and stored as above.

Table 2.3: iSEQ tagged MLST sequencing primers. The polymerase promoters are in lowercase and the loci specific section of the primers are in uppercase.

Primer	Forward/T7 (5' - 3')	Reverse/SP6 (5' - 3')	Product Length (bp)
adhP	cagtaatacgactcactatagggagaaggctGGTGTGTGCCATACTGATTT	cgatttaggtgacactatagaagaggctACAGCAGTCACAACCACTCC	498
atr	cagtaatacgactcactatagggagaaggctATGGTTGAGCCAATTATTTC	cgatttaggtgacactatagaagaggctCCTTGCTCAACAATAATGCC	501
glcK	cagtaatacgactcactatagggagaaggctGGTATCTTGACGCTTGAGGG	cgatttaggtgacactatagaagaggctATCGCTGCTTTAATGGCAGA	459
glnA	cagtaatacgactcactatagggagaaggctAATAAAGCAATGTTTGATGG	cgatttaggtgacactatagaagaggctGCATTGTTCCCTTCATTATC	498
pheS	cagtaatacgactcactatagggagaaggctATATCAACTCAAGAAAAGCT	cgatttaggtgacactatagaagaggctTGATGGAATTGATGGCTATG	501
sdhA	cagtaatacgactcactatagggagaaggctAACATAGCAGAGCTCATGAT	cgatttaggtgacactatagaagaggctGGGACTTCAACTAAACCTGC	549
tkt	cagtaatacgactcactatagggagaaggctACACTTCATGGTGATGGTTG	cgatttaggtgacactatagaagaggctTGACCTAGGTCATGAGCTTT	480

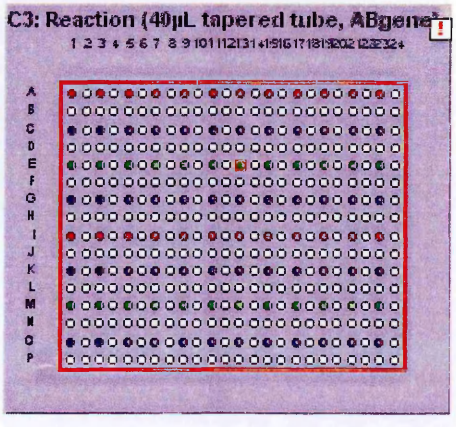
2.1.7.1.2 Reference Database Design

The MLST reference sequences for each of the 7 loci were obtained from pubMLST (101). However, since these reference sequences were generated by Sanger Sequencing the sequence from the end of the reference sequences to the forward and reverse primers was not present, since this is required for iSEQ re-sequencing the missing sequence was filled in with consensus sequence from the whole genome sequence 2603V/R (232) as previously described (91).

2.1.7.2 PCR Amplification

PCR amplification was performed as standard (see 2.1.5) with some modifications. Ninety five 10µl PCR reactions and one negative control were set up in a 384 well PCR plate (ABgene) using the Corbett Robotics CAS4200 (Qiagen) in an interleaved configuration with each sample added sequentially (see figure 2.1). Also, the PCR amplification used the T7/SP6 RNA polymerase promoters (see 2.1.7.1.1). Finally template DNA concentrations where normalised to a constant 5ng/µl. PCR amplification was performed in a 384 well GenAmp 9700 thermal cycler (Life Technologies).

Figure 2.1: PCR plate set-up, coloured wells indicate active wells whereas blank wells are left empty.



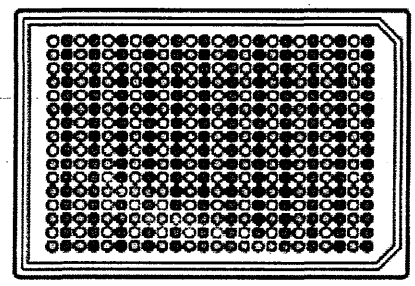
2.1.7.3 Clean-up Using SAP Treatment

Each reaction was treated with 5µl of 0.04U/µl Shrimp Alkaline Phosphatase (SAP) to dephosphorylate any remaining unincorporated nucleotides and render them unavailable in downstream reactions. The SAP solution was added to each well in a Sarstedt 96 well V-bottom plate which was distributed to the active wells in the 384 well PCR reaction plate using the Matrix Plate Mate Liquid Handler (Thermo Fisher Scientific, Hemel Hampstead, UK). Following distribution the PCR plate was sealed and incubated in a 384 well GenAmp 9700 thermal cycler (Life Technologies) at 37°C for 20 minutes followed by 85°C for 10 minutes.

2.1.7.4 Reverse Transcription and Cleavage

Each SAP-treated PCR reaction is split into four 2µl aliquots in a new 384 well PCR plate (ABgene) for the four separate cleavage reactions (C Forward, C Reverse, T Forward and T Reverse) as described in figure 2.2. Each reaction contains 1.08µl RNA free ddH_2O, 0.9µl polymerase buffer, 0.12µl C or T cleavage mixture depending on the reaction, 0.14µl of 100mM DDT, 0.22µl of T7 or SP6 polymerase depending on the reaction and 0.04µl of 0.08 mg/ml RNase A. 4 96 reaction master mixes (CF, CR, TF, TR) were made and each one is distributed to a Sarstedt 96 well V-bottom plate. The Matrix Plate Mate Liquid Handler (Thermo Fisher Scientific) was then used to dispense 2.5µl of cleavage reactions to the appropriate well of the plate which contained 2µl of SAP treated PCR product. The transcription/cleavage reaction plates were sealed and incubated for 3 hours at 37°C for 3 hours in a 384 well GenAmp 9700 thermal cycler.

Figure 2.2: Lay out of a 384 well plate containing reaction mixes for the 4 cleavage reactions and SAP treated PCR product where yellow wells correspond to C Forward, Green to C Reverse, Blue to T Forward and Pink to T Reverse.



2.1.7.5 Conditioning the Plates

Conditioning the reaction mixes de-salts the reactions for use in Mass Spectrometry. The Matrix Plate Mate Liquid Handler (Thermo fisher Scientific, Loughborough, UK) was used to dilute the samples by adding 21.5µl of nano-pure H_2O to each transcription/cleavage reaction. 6 mg of Clean Resin (Sequenom, San Diego CA, USA) was allowed to air dry for 15 minutes before addition to the diluted sample mixes. The plate was sealed and rotated for 10 minutes to agitate the resin in the wells before the plate was centrifuged in a Heraeus Megafuge 11 plate centrifuge (Thermo Fisher Scientific) for 5 minutes at 4000xg for 5 minutes to sink the resin to the bottom of the well .

2.1.7.6 Chip Spotting Using the Nano Dispenser

Conditioned RNA cleavage products and a 4 point size calibrant (Sequenom) were spotted onto SpectroCHIPS (Sequenom) using the NanoDispenser (Samsung, Seoul, South Korea). Seventy microlitres calibrant was added to the calibrant well, the conditioned RNA cleavage products on a 384 well plate were placed into the plate holder at position "MTP-1" and the SpectroCHIP was placed into a chip holder at position at "scout plate 1". The pin head was conditioned in 50% ethanol for 30 minutes before the chip spotting program was ran.

2.1.7.7 Mass Spectrometry of RNA fragments

Using the iSEQ software Plate Editor (Sequenom) the position of each sample, the reference set each sample uses and the reaction transcription/cleavage reaction for each well on the 384 well plate used to spot the chips (see 2.1.7.6) was mapped. This map was exported to the RT workstation which was used to program the MassARRAY compact (Bruker, Billerica MA, USA) using the SpectroACQUIRE software to acquire spectra from the cleavage reactions imprinted onto the SpectroCHIPs (Sequenom). Spectra are obtained in real time and displayed as the MALDI-TOF process progresses.

2.1.7.8 Generating Sequence Data from Spectra

Generated spectra were imported into the iSEQ Analyser software (Sequenom) which was used to calculate the mass of each ionised RNA fragment corresponding to a peak. The size of fragments from which the mass can be inferred reliably is between 4 and 25bp. The mass of each ionised fragment was compared to fragments generated from the in-silico reference sets generated from the reference DNA sequences and the closest matching reference sequence (the one containing the most simulated to observed peak matches) is selected as the closest matching reference sequence. Since the analysis is based on the size of a fragment rather than its composition it is worth considering the probability of a fragment being confused for one of equal mass. Essentially, this can be considered the number of combinations of nucleotides in a given fragment versus the number of permutations. For a single fragment the probability of a fragment matching a different one of the same mass ranged from 13.7% for four nucleotide fragments to $2.9x10^{-10}$ for a 25 nucleotide fragment. However, since the iSEQ system uses 4 reactions the probability of all four reactions identifying the wrong 4 fragments ranges from 0.0003% for four nucleotide fragments to $7.17x10^{-47}$ for a 25 nucleotide fragment making either event unlikely.

The software analysed the closest matching reference sequence compared to the observed peak pattern and assigned probability values referring to the chance of the sequence being correct and to the chance of any variation that has not being observed in the sequence, with a value <0.05 being considered a reliable sequence match.

2.1.8 Sanger Di-deoxy Sequencing

2.1.8.1 Sequencing Primers

All primers were synthesised by Eurofins MWG Operon. Primers are dispatched as lyophilised powder and are made up to a stock concentration of 100pmol/µl and stored at -20°C. For use sequencing primers are made to a concentration of 0.2µM. The sequence of each sequencing primer used is shown in table 2.4

Table 2.4: Forward and reverse sequencing primers

Primer	Scheme	Forward (5' - 3')	Reverse (5' - 3')	Product Length (bp)
adhP Sequencing	MLST	GGTGTGTGCCATACTGATTT	ACAGCAGTCACAACCACTCC	498
atr Sequencing	MLST	ATGGTTGAGCCAATTATTTC	CCTTGCTCAACAATAATGCC	501
glcK Sequencing	MLST	GGTATCTTGACGCTTGAGGG	ATCGCTGCTTTAATGGCAGA	459
glnA Sequencing	MLST	AATAAAGCAATGTTTGATGG	GCATTGTTCCCTTCATTATC	498
pheS Sequencing	MLST	ATATCAACTCAAGAAAAGCT	TGATGGAATTGATGGCTATG	501
sdhA Sequencing	MLST	AACATAGCAGAGCTCATGAT	GGGACTTCAACTAAACCTGC	519
tkt Sequencing	MLST	ACACTTCATGGTGATGGTTG	TGACCTAGGTCATGAGCTTT	480
valS	Coding	ACCAAAATACAATCCTGCCG	AAGCACCCTCAACATCCTTG	581
mmuM	Coding	TCGCTTGACTGTTCAGTTGG	AACAACCTCCGACAACTTGG	629
obgE	Coding	CTAAAGGTATGCACGGTCGAG	TTTTCTTTGAATGCTGCCAA	676
SAG1894	Coding	TGCCAAGACTACTGTTTGGAGA	TCTAGTGCGAAACGATTTTTGA	605
SAG0025	Coding	CTACTACATTCCAGGTCAAGC	GAAGATGATATTTGCCTTAGC	537
SAG0027	Coding	TGTGTTGCCATGTGTGTCAA	ATTTCAAACATTTCTTCGTGC	590
SAG0043	Coding	TAAAGCCTACATCGAAGAGCA	CTTTCTCGTAATCGAGCGGAT	601
SAG0047	Coding	GTACCCACGGTGTTCACGC	CCAAAAGTTGATTCCATATTGCGC	590
cpsL	Coding	GCACCATTAGTTGGTTTTTCT	TCTCCTCCCATTATTTGAGC	626
cylB	Coding	AGGTGCCTTTGGAGTTATGG	TAGAGACAGTGGCTTCGTTGG	500
SAG0032 Non-Coding	Non-Coding	TTGGAATGCAATGCCAGAT	GGCATAACACCTCCACTTTCA	435
SAG0043 Non-Coding	Non-Coding	GCAGGGCAGGACAAAATCTA	TGGGCAGAGACGACTTTCTT	520
SAG0106 Non-Coding	Non-Coding	ACGAAGAAGACGATGAAGAGGA	ACCTTTAGCCCACGGTTTTT	420
SAG0649 Non-Coding	Non-Coding	CCTAAAACAGGGGGAATTGG	CCGTTGCAGATTGTCCTCTT	469
SAG1768 Non-Coding	Non-Coding	TGTGGTTTGACTGACATGCTG	TCAAACTGTTGACGGTAACCAA	402
SAG2143 Non-Coding	Non-Coding	ACCTTTAGCAGAGCCACCTAT	TCACCAAAATGATGGGACTT	431
SAK_1320 Non-Coding	Non-Coding	AACTGACATGCCTTGCGTAG	TGACCACTTTCTTCTTGGGATT	411

2.1.8.2 PCR Product Clean-up Using AMPure

PCR product clean up was performed using AMPure (Beckman Coulter, High Wycombe, UK) reagents either manually or automated using the Biomek NX^P (Beckman Coulter) laboratory automation workstation. Both methods are detailed below.

To perform PCR product clean up manually first PCR product was transferred to a MicroAmp optical 96 well reaction plate (Life Technologies) and 45µl of AMPure beads were added to each 25µl reaction, mixed by pipetting then incubated for 5 minutes to allow PCR product to bind to the magnetic beads. The PCR plate was then placed onto a magnetic SPRIPlate (Beckman Coulter) and incubated until magnetic beads were visibly bound to the outside of the PCR plate (approximately 5-10 minutes). Cleared solution from the PCR plate was aspirated and discarded and two 200µl 70% ethanol washes were performed. After the second ethanol aspiration the PCR plate was air dried for 20 minutes. The beads were then eluted in 40µl dH_2O and mixed by pipetting 10 times to ensure PCR product released from the beads. Whilst on the magnetic plate, 35µl of the eluted product were transferred to a fresh 96 well MicroAmp optical reaction plate (Life Technologies) for sequencing.

Automated PCR product clean up uses the same chemistry on the Biomek NX^P (Beckman Coulter) laboratory automation workstation with two changes to protocol incubation times. Firstly, the time to bind beads to the side of the well after AMPure addition was seven and a half minutes (opposed to between 5-10 minutes with visual verification of bead binding). Secondly, after the ethanol washes the plate is dried for 15 minutes opposed to 20 in the manual reaction. Programs for performing AMPure clean up were supplied with the equipment by the manufacturer and controlled using the Biomek software (Beckman Coulter).

2.1.8.3 Cycle Sequencing

Cycle sequencing used BigDye v1.1 and v3.1 (Life Technologies) chemistry were used, the reaction mixes were the same for both versions. Cycle sequencing plates were set up using the QIAgility PCR robotics platform (Qiagen). Each reaction contained 1µl BigDye (Life Technologies), 1.5µl 5x Sequencing Buffer (Life Technologies), 1µl Sequencing Primer (0.2µM), 3µl AMPure cleaned up PCR product (containing between 5 and 20ng DNA) and 3.5µl molecular biology grade H_2O (Fisher Scientific). Cycle sequencing reactions were performed using either a GenAmp 9700 96 well thermal cycler or a Veriti 96 well cycler (Life Technologies). Cycling conditions were an initial denaturation of 1 min at 96°C followed by 25 cycles of a 10 second denaturation at 96°C, a 5 second annealing at 55°C and a 4 minute extension at 60°C for the GenAmp 9700 thermal cycler, the Veriti thermal cycler uses the same conditions except for a shorter extension time of 1 minute 15 seconds.

2.1.8.4 Cycle Sequencing Product Clean-up Using CleanSEQ

Cycle sequencing product clean up was performed both manually and automated using the Biomek NX^P (Beckman Coulter) laboratory automation workstation.

Manual cycle sequencing product clean up was performed using cleanSEQ (Beckman Coulter). 10µl of cleanSEQ beads and 42µl of 85% ethanol were added to each 10µl cycle sequencing reaction in a MicroAmp optical 96 well reaction plate (ABI). The mixture was pipette mixed until the solution was homogenous and the reaction place was placed onto a SPRIPlate (Beckman Coulter) until magnetic beads were visibly bound to the outside of the PCR plate (approximately 3-5 minutes). The cleared solution from the PCR plate was aspirated and discarded and two 85% ethanol washes were performed. The plate was then left to air dry for 10 minutes and the beads were eluted in either formamide or molecular biology H_2O depending on the length of time from clean up to sequencing, i.e. clean up performed the same day was sequencing was eluted in water, for a longer time between clean up and sequencing formamide was used. The eluted

product was then placed onto a SPRIPlate (Beckman Coulter) until the beads were bound to the side of the PCR plate and 35µl was removed and added to a clean PCR plate for sequencing.

Automated cycle sequencing product clean up uses the same chemistry on the Biomek NX^{P} (Beckman Coulter) laboratory automation workstation. One change to the incubation times was made from the manual reaction and the beads where left for three minutes (opposed to between three and five minutes with visual verification). Programs for performing cleanSEQ clean up were supplied with the equipment by the manufacturer and controlled using the Biomek software (Beckman Coulter).

2.1.8.5 Capillary Electrophoresis Using ABI 3130xl/3730 Sequencers

Sequencing was performed by capillary electrophoresis using either the 3130xl or 3730 genetic analyser (Life Technologies). Either 35µl of cleaned up cycle sequencing product without magnetic beads in a 96 well MicroAmp optical reaction plate (Life Technologies) were placed in a plate retainer or 80µl of cleaned up cycle sequencing product with magnetic beads, also in a 96 well MicroAmp optical reaction plate (Life Technologies) were placed onto a direct inject plate (ABI) and assembled with a plate retainer. Sequencing was performed using a 3 second injection time and the standard Genomic Sequencing Service (GSU) analysis protocol specific to either V1.1 or v3.1 BigDye (Life Technologies) chemistry was used to base-call and trim sequence using a rule of >4 bases out of 20 had to have a quality value (QV) >20.

2.1.8.6 Trace Assessment and Assembly

Trace files were checked using Sequence Scanner v1.0 (Life Technologies) to assess trace quality by ensuring the quality values (a numerical value between 1-100 assigned to each base in chromatogram that shows the level of confidence in that base) over the length of the sequence are >20 and by comparing the experimentally determined contiguous read length to the expected read length. Forward and reverse reactions were assembled into contigs using SeqMan in the Lasergene 8 software package (DNASTAR, Madison, WI, USA) where sequences are trimmed down to good quality bases and the forward and reverse strand are checked for conflicts and manually base called.

2.2 Bioinformatic Methods

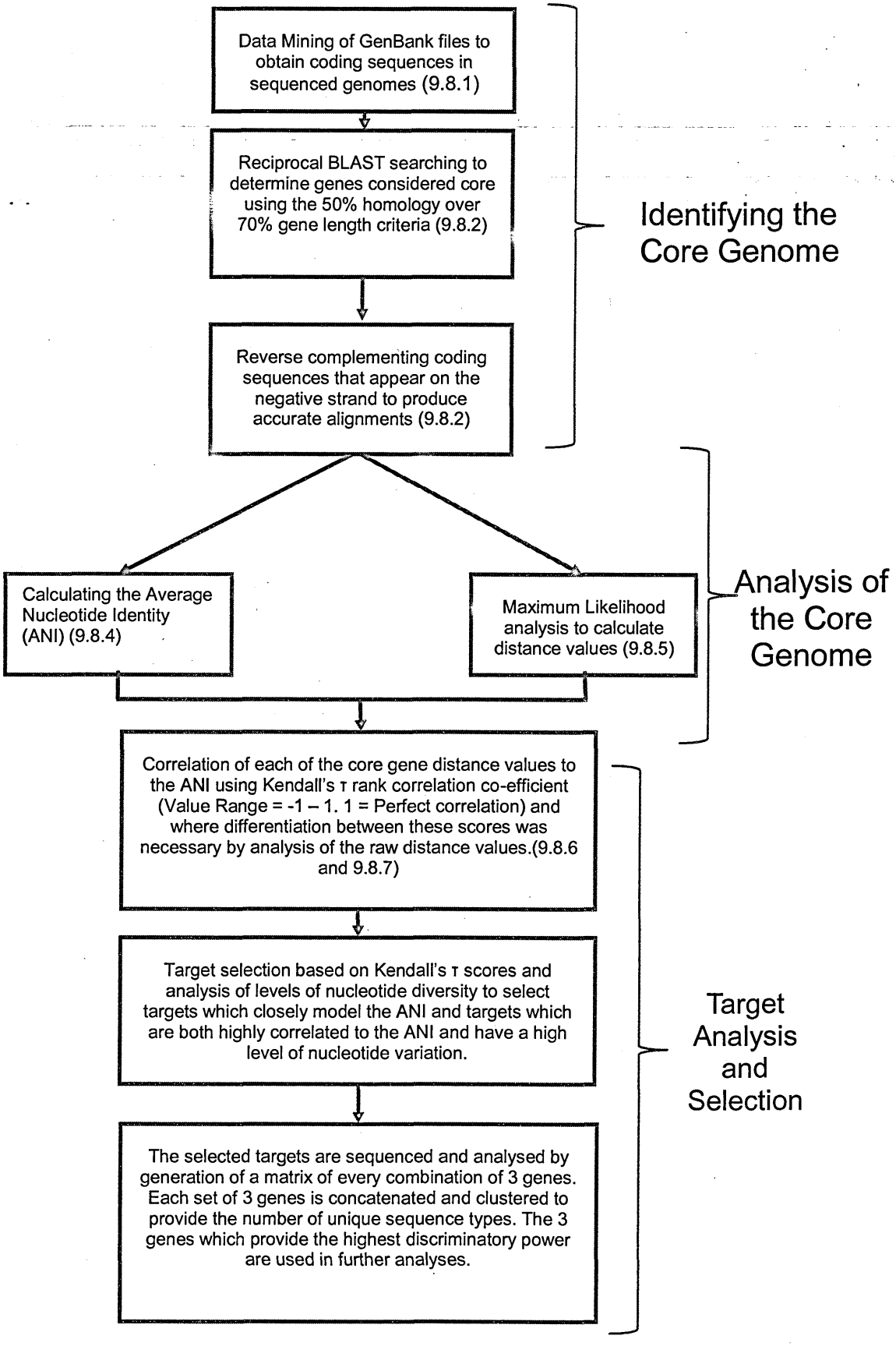
2.2.1 Sequenced Genomes Used

Three fully sequenced *S. agalactiae* genomes, serotype V 2603V/R (232), serotype III NEM316 (68) and serotype IA A909 (231) as well as five whole genome shotgun sequences 18RS21, 515, CJB111, COH1 and H36B (231) with the serotypes II, IA, V, III and IB respectively were obtained from GenBank (http://www.ncbi.nlm.nih.gov/genomes/lproks.cgi) for use in this study.

2.2.2 Core Genome Analysis

The core genome was extrapolated and analysed as described by Konstantinidis (119) by identifying the core genome using reciprocal BLAST, performing maximum likelihood analysis of each core gene, determining the Average Nucleotide Identity (ANI) of the genome and selecting profiling targets by correlating each core gene to the ANI using a number of methods. The overview of this process is shown in figure 2.3

Figure 2.3: Overview of the target selection process



2.2.2.1 Data-Mining GenBank Files for Coding Sequences

Coding sequences from the GenBank files of the three fully sequenced genomes were extracted using a custom PERL script adapted from a previous study (1) which used the PERL modules SeqIO and AnnotationCollectionI to identify coding sequences (appendix 9.8.1), and their locus tag and output these sequences in FASTA format. For the five whole genome shotgun genomes the cDNA regions were taken directly from NCBI in FASTA format and custom PERL scripts were used to split each coding region into a separate file and change the NCBI reference identifiers to the assigned locus tags.

2.2.2.2 Reciprocal BLAST

Reciprocal BLAST was used to identify the core genome of 1) the three fully sequenced *S. agalactiae* genomes and 2) all 8 *S. agalactiae* genomes. In both cases all coding sequences were concatenated into a singe Fasta file and BLAST databases were constructed using formatdb including the –o option to generate additional indexing files, a custom PERL script from a previous study (appendix 9.8.2) was then adapted to perform reciprocal BLAST using the genome 2603V/R as a reference genome. Each gene from the 2603V/R genome was BLASTed against each of the test genomes and a gene was considered core if it had hits in all other genomes that were 50% homologous over 70% of the gene length and each hit would meet the same criteria when BLASTed back against the reference genome.

2.2.2.3 Determining the COG Categories of Core Genes

Cluster of Orthologous (COG) categories for each gene in the reference genome 2603V/R were taken from NCBI (http://www.ncbi.nlm.nih.gov/sutils/coxik.cgi?gi=252) and concatenated into a single Excel worksheet, core gene locus tags were added and the VLOOKUP function was used to identify the core genes in the list of all COG categorised genes and therefore isolate COG categories of each core gene.

2.2.2.4 Alternate Methods for Determining the Core Genome

Two alternate methods where used to identify the core genome to confirm reciprocal BLAST results. Firstly, the Multi-Genome Homology Comparison tool (JCVI, http://cmr.jcvi.org/cgi-bin/CMR/shared/MakeFrontPages.cgi?page=circular_display) which used protein sequences to determine homology. A 50% sequence identity criteria was used to identify the core genome of both the three fully sequenced genomes and of all eight *S. agalactiae* genomes. Secondly the Panseq online tool (124) was used to identify the core genome of the three fully sequenced strains. However, the output of this is a core genome alignment, so it identifies potential coding sequences and the GLIMMER software (40,201) was then used to identify the genes within one core genome alignment ortholog.

2.2.2.5 Reverse Complementing Negative Strand Genes

Core genes are often found on different strands and therefore could not be aligned accurately without ensuring all genes were in the correct orientation, the fully sequenced genomes were all orientated correctly however the whole genome shotgun genes, because of the way in which they were extracted needed to be manually checked using BioEdit and re-aligned if necessary.

2.2.2.6 Alignment Using ClustalW

A custom PERL script adapted from a previous study (appendix 9.8.3) was used to align each set of core gene orthologs using the ClustalW PERL module and output the alignment in Nexus file format for use with the PAUP* program for maximum likelihood analysis.

2.2.2.7 Calculating the Average Nucleotide Identity (ANI)

The ANI of the core genome was calculated using a custom PERL script which used AlignIO (appendix 9.8.4) to calculate the number of identical bases between each ortholog pair for each core gene. The ANI between each ortholog pair was then calculated using the equation below.

$$ANI = \frac{\Sigma\left(\frac{Number\ of\ Identical\ Bases}{Alignment\ Length}\right)}{Number\ of\ Core\ Genes}$$

The output was a file containing the average nucleotide identity between each pair of genomes and the ANI was calculated by taking the average of these values.

2.2.2.8 Maximum Likelihood Analysis of Each Core Gene

A custom PERL script (appendix 9.8.5) was used to perform maximum likelihood analysis on each core gene by taking each gene and using PAUP* (254) to generate a file containing the information necessary for ModelTest v3.7(182) (Gamma Shape Parameter, Proportion of Invariable Sites, Base Pair Frequency, ti/tv ratios and log likelihood values). This information was then used to analyse each gene against 56 models of evolution to determine the most likely model using the Akaike information criterion. The best model for each individual core gene is then inputted back into PAUP* using the PAUP block to perform maximum likelihood analysis using the most likely model of evolution. The output of this is a set of files for each core gene which contains the evolutionary distance values between each ortholog of each core gene.

2.2.2.9 Target Selection

2.2.2.9.1 Kendal's Rank Correlation Coefficient

The Kendall rank correlation coefficient (also known as the Kendall's tau (τ) coefficient) is a statistic used to measure the correlation between two measured quantities. It is a non-parametric hypothesis test for statistical dependence which measures the similarity of the order of given data (i.e. a rank correlation).

The evolutionary distance values were formatted for use in the Stata program (StataCorp) using custom PERL scripts (appendix 9.8.6 and 9.8.7) to add the genome distance values (ANI) to each core gene file and changing the file name to an ordered numerical value and creating a file key (the new file number and its corresponding locus tag). These files were then analysed using a Stata script to measure the TauA and TauB statistics between each core gene and the ANI.

2.2.2.9.2 Absolute Subtraction

The Kendal's Tau test was used to identify genes that are closely related to the whole genome average. However, using the three fully sequenced genomes Kendal's Tau gave insufficient discriminatory power and so the Absolute Subtraction method was devised. The distance values of the ANI and each core genome ortholog were totalled using a custom PERL script (appendix 9.8.8) and in Excel the total core genome ortholog distance value was absolutely subtracted (i.e. negative results were converted to a positive result) and therefore Absolute Subtraction scores close to zero are more closely related to the ANI.

2.2.2.9.3 Analysis of the Average Number of SNP's

Selected core gene orthologs were aligned in BioEdit and a distance matrix was created. The distance matrix was used to calculate the Average Number of SNP's per ortholog and per 100bp per ortholog using Excel.

2.2.2.9.4 Clustering of Sequence Data

Sequence data was analysed using CD-Hit est software (135) to determine the number of unique allele types in a given set of sequences using a homology criteria of 1.00 (100%) and a word size of 9. Concatenated DNA sequences of multiple DNA targets were also analysed using CD-Hit est using the same criteria. Here a custom PERL script was used to automate the use of the script fastaConcat.pl, written by D. Wolf and N. Takebayashi (Institute of Arctic Biology) and downloaded from (http://hi.baidu.com/cnelon1133/blog/item/ab545df11883eacf7931aabb.html accessed March 2008), to concatenate every combination of sequenced DNA from selected loci, input each sequence into CD-Hit and return an output of the combination of loci and the number of unique sequence types.

2.2.2.10 Analysis of Sequence Data

2.2.2.10.1 jModelTest

jModelTest (181) was used for sequence data analysis in place of ModelTest v3.7 (182) since it is independent of PAUP*, uses an intuitive graphical user interface (GUI) and was shown to be faster. Phylogenetic model selection was performed by inputting Fasta sequence files and selecting "compute likelihood scores" (which performs the same function as PAUP* did for ModelTest) using the default settings. Following computation of likelihood scores AIC calculations are performed to select the most likely model of evolution for use in phyML.

2.2.2.10.2 phyML

DNA sequences were converted from FASTA into phylip format using Seaview v4 (70) and entered into phyML (72). Trees were constructed using either a default model (GTR with optimised gamma shape parameter over 4 sites, optimised proportion of invariable sites and empirical base frequencies) or using the model selected using JModelTest. The selected model was inputted by selecting custom model and inputting the correct substitution code (for example, phyML does not

have the TIM3ef model whose substitution code is 012032), inputting the jModelTest selected ti/tv rates which is the ratio of the number of transition (purine-purine or pyrimidine-pyrimidine) to transversion (purine- pyrimidine or vice versa) substitutions that appear to have occurred since two sequences separated from a common ancestor and base frequencies, if necessary specifying the gamma shape parameter and proportion of invariable sites. All trees were bootstrapped 100 times.

2.2.2.10.3 BioNJ

BioNJ (67) was used to create distance trees using the ANI distance values. BioNJ uses a neighbor joining algorithm which is a bottom-up clustering method for the creation of phylogenetic trees created by Saitou and Nei (200). The method works by creating a "Q matrix" from the distance values and joining the two most closely related taxa based on these scores. The distance of the new node created from these joined taxa to the rest of the tree is then calculated. This process is repeated using the new node and the closest taxa to it as a new pair until the tree is fully resolved. Here the ANI script output was converted into phylip format and inputted into the BioNJ program and trees were constructed using the default settings (i.e. A Jukes-Cantor substitution model, a Gamma Distribution Parameter of 1 and a ti/tv ratio of 2).

2.2.2.10.4 pubMLST

Assignment of MLST allele types and sequence types was performed using the *S. agalactiae* database on the pubMLST website (http://pubmlst.org/sagalactiae/) (101). Fasta formatted DNA sequences were aligned against allele type 1 for all loci and trimmed to the correct length and inputted as a "single locus batch query". Results for each loci are saved and stored in an excel worksheet. Complete sets of allele types for each strain are then entered as a "batch profile query" to obtain the sequence type for each strain.

2.2.2.10.5 START2

START2 (Sequence Type Analysis and Recombinational Tests Version 2) (102) was used to measure Linkage Disequilibrium between allele types of each combination of profiling loci. Each profiling scheme was also tested for potential recombination sites between alleles of each sequenced loci using a Maximum Chi squared test.

To calculate linkage disequilibrium, for each profiling scheme the complete collection of allelic profiles/ST assignments and DNA sequences of each allele type were inputted into the software. Two separate Index of Association tests were performed, a "Classical" (Maynard Smith) and a "Standardized" (Haubold) test was performed and it was considered evidence for/against linkage disequilibrium if both tests were in agreement (102).

The Maximum Chi squared test was performed by inputting sequence data as above and running a batch Maximum Chi squared test to identify potential recombination sites between every combination of alleles.

2.2.3 Analysis of Mono-Nucleotide Repeats (MNRs)

2.2.3.1 Identification of Virulence Factors

Virulence factors of the three fully sequenced *S. agalactiae* genomes were obtained from VFDB (Virulence Factors Database) (31). To obtain virulence factors for the partially sequenced genomes each virulence factor from the genome 2603V/R was BLASTed against a database of all coding sequences from the 5 wgs genomes and homologues matching with an e value of $> 10^5$ were considered homologues and therefore virulence factors. Selected loci from A909 and NEM316 were also blasted against the wgs coding sequence database if these loci did not have a homologue in 2603V/R.

2.2.3.2 Location of MNRs

MNRs were identified in non-coding and coding DNA regions as potential profiling markers. Analysis of non-coding DNA was performed using the three fully sequenced genomes and analysis of coding DNA was performed using the protein coding DNA sequences from all eight genomes.

Non-coding DNA was investigated for MNR tracts using custom PERL scripts (appendix 9.9.1, 9.9.2 and 9.9.3) designed to separate the coding sequences on the positive and negative strand of the chromosome, format the coding sequences into tab delimitated format containing the locus tag and the sequence of each gene on a separate line of a new file. These files were used to replace the coding DNA minus 20bp at each end of the coding gene (to overcome overlapping coding genes not being replaced and as a site for primer design) in the sequenced genome file with the appropriate locus tag, meaning the remaining DNA was either non-coding or coding for a protein on the opposite strand (this problem was solved by only using non-coding DNA between 200-300bp and manually checking selected markers using BLAST). Each region of DNA between 2 locus tags was cut out and stored to a separate file with each non-coding region on a separate line along with the locus tags at the 5' and 3' ends of the DNA. Each non-coding region was then used in a custom designed MNR finder script which returns the 5' and 3' locus tags, the length and base composition of any MNRs found and the length of the non-coding region.

Coding DNA was analysed by modifying the custom MNR script above (appendix 9.9.4) to analyse the tab delimitated coding sequences (also above) and identify MNRs in coding DNA and return the locus tag, any repeats found, the position of the start of the repeat and the length of the sequence. Another custom PERL script was developed to locate any of the 3 stop codons (TAA, TAG and TGA) in all open reading frames (ORFs) and return the locus tag, position and which stop codon was located.

2.2.3.3 Identification of Genes Containing Homopolymeric tracts

Previous work has shown that homopolymeric tracts represent a general regulatory mechanism in prokaryotes (171) and the most reliable indicator or a MNR tract that may be regulated by slipped strand mispairing is a repeat occurring in the first 10% of the gene length. The tab delimitated file from above was imported into an excel worksheet and the position of the repeat was calculated as a percentage of the total gene length for each gene. Each list of genes was correlated to known virulence factors and any virulence factor with repeats in the first 10% of the gene was considered to potentially be regulated by slipped strand mispairing.

2.2.3.4 Identification of Non-coding Regions for Profiling

The locus tags at the 5' and 3' ends of non-coding regions of between 200 and 300bp which contained MNR repeat tracts were correlated to known virulence factors gained from VFDB (31) using Excel. Non-coding regions with a virulence factor at either the 5' and/or 3' of the gene were considered suitable for sequence analysis (see above). Additionally, three random non-coding DNA sequences between 200-300bp in length that had no MNRs and were not associated with any virulence factors were selected as controls.

Chapter 3

MLST Profiling

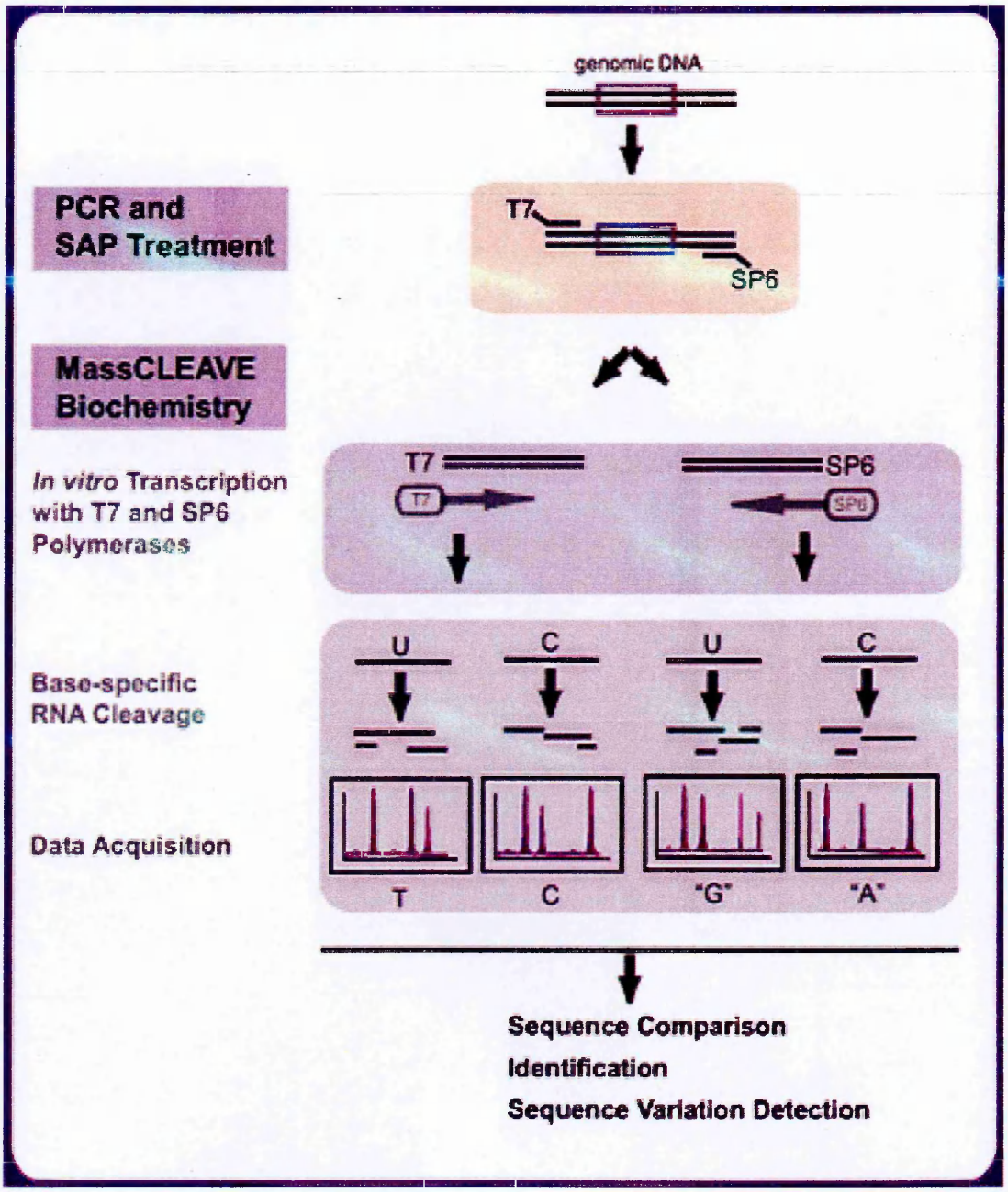
3.0 MLST Profiling

3.1 Introduction

MLST is currently the gold standard technique for typing GBS (103). This was performed on the strain collection of 135 clinical and reference strains to determine whether the Sequenom iSEQ platform was suitable for re-sequencing GBS and to provide comparative data for novel methods that were developed.

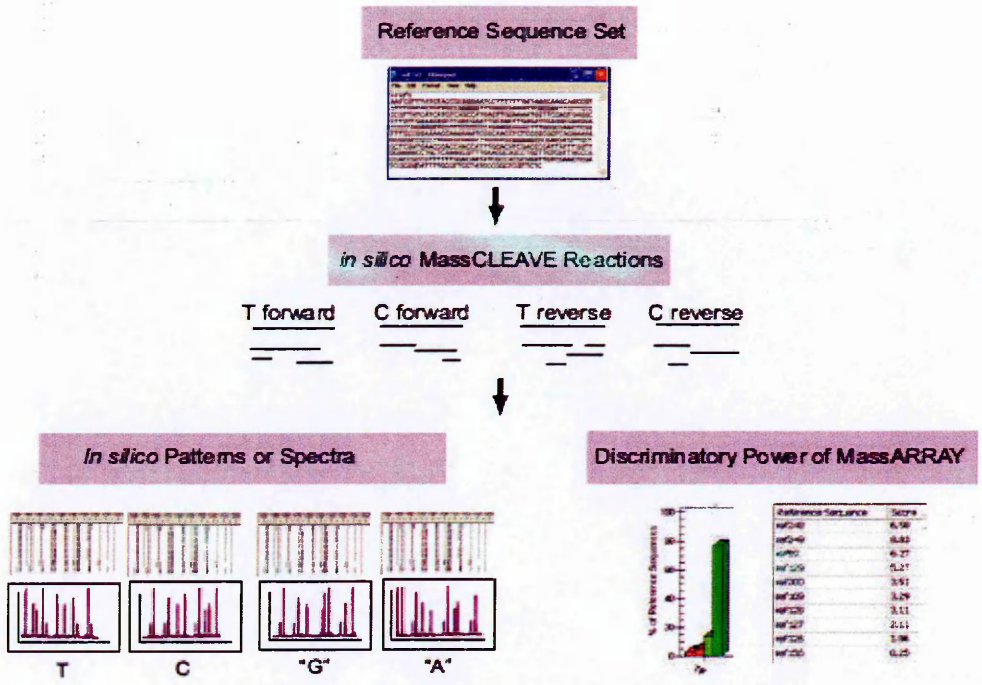
The Sequenom iSEQ platform is a re-sequencing technology that uses Matrix Assisted Laser Deabsorption/Ionisation Time of Flight Mass Spectrometry (MALDI-TOF MS). In this method PCR product reverse transcribed into RNA and cleaved in four separate reactions at T and C bases on the forward and reverse strand (figure 3.1) and the mass of each cleavage fragment is measured using MALDI-TOF MS. MALDI-TOF MS measures the mass of fragments by directing a laser pulse at a chemical matrix molecule which absorb UV from the laser leading to the ablation of the upper layer of the matrix material which produces a hot plume of matrix which transfers charge to the RNA fragments. The ionised RNA fragments are then carried through a vacuum by an electromagnetic field to a detector. The time it takes the ionised RNA fragments to reach the detector is used to calculate their mass (87). The fragment sizes are then compared to a simulated cleavage pattern (figure 3.2) from a user generated in-silico reference set, the more observed fragments matching the simulated fragments the higher the probability of a confident match. The iSEQ platform has previously been demonstrated as suitable for MLST on *Neisseria meningitidis* (91,109) and should therefore be suitable for application to the GBS MLST scheme.

Figure 3.1: The iSEQ experimental process



Source: iSEQ user manual

Figure 3.2: Generating simulated fragment patterns from a user supplied reference set



Source: iSEQ user manual

3.1.1 Assessing the iSEQ Platform for MLST Profiling

Initially, to test the system and to get an indication of failure rates, each MLST loci was sequenced for six clinical isolates by both Sanger sequencing and using the iSEQ platform. For iSEQ sequencing the MLST sequencing primers (103) were tagged with the T7 and SP6 tags and DNA was amplified, reverse transcribed and cleaved at T and C bases on the forward and reverse strands in four separate reactions before being conditioned and spotted to spectroCHIPS for analysis by MALDI-TOF MS. Sanger sequencing was performed as previously described except the sequencing primers were also used for amplification (103). Using the iSEQ platform sequences were generated for 35/42 targets (83%) and for Sanger sequencing sequences were generated for 34/42 targets (80%), showing a similar failure rate between the two methods. The targets that failed to sequence were repeated and sequences from each method were compared showing that

the sequences generated from the iSEQ platform were identical to the sequences identified using Sanger sequencing. This initial assessment suggest that the platform is suitable for MLST profiling of GBS.

3.1.2 Optimisation of iSEQ for MLST Profiling

The 7 loci of the MLST scheme were re-sequenced for 135 isolates using the Sequenom iSEQ system and profiles were obtained for 134 (one isolate was a mixed sample and a profile could not be obtained) and were matched to existing sequence types taken from pubMLST (101). Each sequence match is placed into one of three categories, a confident match, a sequence that shows a significant probability (>0.05) of sequence variation from the closest reference sequence or a bad assay indicating the reaction has either failed or the sequence is so divergent that the sequence cannot be matched to any reference sequence.

The first step was to determine if the iSEQ results were correct. Sequenom states that a confident match is a match to a reference sequence with a sequence variation probability less than 0.05, however over half of the sequences generated (55%) were shown to have a significant chance of sequence variability. For the 7% of sequences that failed after being sequenced twice using iSEQ Sanger sequencing was performed (table 3.1).

Table 3.1: The number of sequences per loci falling into each of the three match categories

Loci	Confident Matches	Probability of Sequence Variation >0.05	Sanger Sequenced	Other Information
adhP	46	72	16	
atr	6	120	8	
glcK	59	70	1	4 Loci contained 2KB inserts
glnA	96	34	4	
pheS	19	112	3	
sdhA	106	6	22	
tkt	16	104	14	

From the 55% of sequences showing significant chance of sequence variability it is unlikely that they are not present in the MLST database, especially since in the MLST profiles database the vast majority of profiles are made up of allele types discovered in the original MLST paper (103). For example, the prevalence of allele type 1 in allelic profiles for each loci ranges from 20% (*tkt*) to 71% (*pheS*) suggesting that allele types discovered early on account for a high proportion of allele types. It was therefore necessary to determine why such a large amount of sequences were falling into this category.

Loci showing more variation should be easier to discriminate using the iSEQ platform. Thus the level of nucleotide variation was measured between all identified sequence types of all loci of the GBS MLST scheme. They were compared to the same information from the successful *N. meningitidis* MLST scheme (91). This showed significantly higher levels of average nucleotide variation between all loci of the *N. meningitidis* MLST scheme (tables 3.2 and 3.3). The GBS loci ranged from 0.65-1.04 SNPs per 100bp wheras the *N. meningitidis* ranged from 2.7-11.84 SNPs per 100bp. So even the lowest variation observed in a *N. meningitidis* loci is significantly higher than the GBS loci and this may be contributing to the low level of confidence in the GBS iSEQ results.

The average AT content of all previously identified sequence types from all loci of both the GBS and *N. meningitidis* MLST schemes were also compared. T bases do not fly well in a mass spectrometer due to T bases not accepting charge and therefore fragments containing an abundance of T bases may not be identified by MALDI-TOF MS (C. Honisch, Personal Correspondence). This could lead to incorrect sequence type assignments or affect the statistics relating to the confidence of a match. This analysis showed that the AT content was significantly lower in the loci of the *N. meningitidis* MLST scheme (table 3.4)

Table 3.2: The number of reference sequences, length of loci and SNP's per sequence/per 100bp for the GBS reference set

Locus	Sequence Types	Length	Average SNP's	SNP/100bp
adhP	78*	498	4.46*	0.9
atr	53	501	5.21	1.04
glcK	39	459	3.77	0.82
glnA	46	498	3.32	0.67
pheS	31	501	3.25	0.65
sdhA	40	519	5.25	1.01
Tkt	33	480	4.2	0.88

**adhP* allele types 49 + 54 removed due to a 163bp deletion which would disproportionately bias the SNP scores

Table 3.3: The number of reference sequences, length of loci and SNP's per sequence/per 100bp for the N. meningitidis reference set.

Locus	Sequence Types	Length	Average SNP's	SNP/100bp
adcZ	454	433	29.14	6.73
adk	307	465	18.23	3.92
aroE	514	490	58.02	11.84
fumC	461	465	12.56	2.7
gdh	496	501	31.9	6.37
pdhC	478	480	26.42	5.5
pgm	474	450	30.25	6.72

Table 3.4: The AT content of each target of the GBS MLST scheme compared to the AT content of the N. meningitidis MLST scheme loci

GBS	
Loci	AT Content
adhP	56.22
atr	62.67
glcK	57.52
glnA	64.06
pheS	62.87
sdhA	58.58
tkt	60.83

N. meningitidis	
Loci	AT Content
adcZ	48.5
adk	47.74
aroE	42.66
fumC	42.37
gdh	48.1
pdhC	46.46
pgm	45.56

All targets in the *N. meningitidis* MLST scheme are significantly more variable than the targets of the GBS MLST set and the AT content is on average 15% higher in the GBS MLST set. The combination of this data suggests a possible explanation as to why the GBS scheme is not performing as well on the iSEQ platform as the published *N. meningitidis* MLST scheme. Initial

testing did show that the sequences generated by the iSEQ platform matched sequences generated by Sanger sequencing. Since that used only a limited number of isolates it is still possible that the sequences were being incorrectly identified or that the statistics the software uses were not suited to these loci despite the sequences being identified correctly. To investigate further a selection of 30 isolates where Sanger sequenced for the *atr*, *pheS* and *tkt* loci, the loci which demonstrated the highest number of matches with significant sequence variation probabilities. The results are shown in tables 3.5, 3.6 and 3.7.

Table 3.5: Comparison of iSEQ assigned sequence types to Sanger sequencing for the atr loci

Strain	Sequenom ST	Confidence	Score	P Variation	Match?
0933	*atr_-6	Sequence variations	0.921	0.219	Yes
8186	*atr_-6	Sequence variations	0.938	0.063	Yes
8188	atr_-1	Sequence variations	0.879	0.894	Yes
00/46	*atr_-2	Sequence variations	0.934	0.079	Yes
00/465	*atr_-2	Sequence variations	0.921	0.227	Yes
03/0477/GB	*atr_-6	Sequence variations	0.913	0.342	Yes
03/226	*atr_-3	Sequence variations	0.915	0.295	Yes
03/270	*atr_-3	Sequence variations	0.895	0.685	Yes
03/414	atr_-4	Sequence variations	0.887	0.973	No
03/427	atr_-4	Sequence variations	0.914	0.753	Yes
03/438	atr_-2	Sequence variations	0.925	0.147	Yes
03/439	*atr_-3	Sequence variations	0.936	0.048	Yes
03/451	*atr_-6	Sequence variations	0.915	0.295	Yes
03/460	atr_-4	Sequence variations	0.9	0.916	Yes
03/464	atr_-1	Sequence variations	0.914	0.316	Yes
03/467	*atr_-2	Sequence variations	0.894	0.689	Yes
03/471	*atr_-2	Sequence variations	0.915	0.303	Yes
03/474	*atr_-3	Sequence variations	0.921	0.212	No
03/478/GB	*atr_-3	Sequence variations	0.909	0.402	Yes
515	*atr_-6	Sequence variations	0.929	0.12	Yes
GBS III	atr_-1	Sequence variations	0.933	0.106	Yes
GBS IV	atr_-4	Sequence variations	0.911	0.819	Yes
GBS VII	*atr_-2	Sequence variations	0.926	0.174	Yes
GBS VIII	*atr_-2	Sequence variations	0.926	0.168	Yes
GBS X	*atr_-1	Sequence variations	0.856	0.996	Yes
H03420043	atr_-1	Sequence variations	0.935	0.085	Yes
H034540119	atr_-4	Sequence variations	0.905	0.934	Yes
H035140030	atr_-4	Sequence variations	0.916	0.654	Yes
H040200291	*atr_-3	Sequence variations	0.913	0.337	Yes
H040540417	atr_-1	Sequence variations	0.923	0.209	Yes

Table 3.6: Comparison of iSEQ assigned sequence types to Sanger sequencing for the pheS loci

Sample	Sequenom ST	Confidence	Score	P Variation	Match?
0933	pheS-4	Sequence variations	0.93	0.109	Yes
8186	pheS-4	Sequence variations	0.928	0.202	Yes
8188	*pheS-1	Sequence variations	0.857	0.997	Yes
00/46	*pheS-1	Sequence variations	0.917	0.259	Yes
00/465	*pheS-1	Sequence variations	0.918	0.257	Yes
03/0477/GB	pheS-4	Sequence variations	0.936	0.075	Yes
03/226	*pheS-1	Sequence variations	0.939	0.328	Yes
03/270	pheS-1	Sequence variations	0.937	0.061	Yes
03/414	*pheS-1	Sequence variations	0.902	0.513	Yes
03/427	*pheS-1	Sequence variations	0.926	0.158	Yes
03/438	*pheS-1	Sequence variations	0.93	0.12	Yes
03/439	*pheS-1	Sequence variations	0.937	0.089	Yes
03/451	pheS-4	Sequence variations	0.912	0.346	Yes
03/460	*pheS-1	Sequence variations	0.923	0.203	Yes
03/464	*pheS-1	Sequence variations	0.879	0.876	Yes
03/467	*pheS-1	Sequence variations	0.917	0.264	Yes
03/471	*pheS-1	Sequence variations	0.92	0.204	Yes
03/474	*pheS-1	Sequence variations	0.909	0.401	Yes
03/478/GB	*pheS-1	Sequence variations	0.928	0.134	Yes
BAA1177(515)	*pheS-4	Sequence variations	0.942	0.067	Yes
GBS III	*pheS-1	Sequence variations	0.934	0.134	Yes
GBS IV	pheS-1	Sequence variations	0.946	0.054	Yes
GBS VII	pheS-1	Sequence variations	0.941	0.083	Yes
GBS VIII	pheS-1	Match	0.957	0.022	Yes
GBS X	*pheS-1	Sequence variations	0.891	0.74	Yes
H03420043	pheS-1	Sequence variations	0.942	0.11	Yes
H034540119	*pheS-4	Sequence variations	0.912	0.34	Yes
H035140030	*pheS-1	Sequence variations	0.936	0.09	Yes
H040200291	*pheS-1	Sequence variations	0.938	0.067	Yes
H040540417	pheS-3	Sequence variations	0.925	0.162	Yes

Table 3.7: Comparison of iSEQ assigned sequence types to Sanger sequencing for the tkt loci

Sample	Sequenom ST	Confidence	Score	P Variation	Match?
0933	*tkt_-3	Sequence variations	0.873	0.749	Yes
8186	tkt_-3	Sequence variations	0.957	0.053	Yes
8188	tkt_-1	Sequence variations	0.941	0.49	Yes
00/46	*tkt_-2	Sequence variations	0.906	0.207	Yes
00/465	*tkt_-2	Sequence variations	0.973	0.019	Yes
03/0477/GB	*tkt_-3	Sequence variations	0.895	0.372	Yes
03/226	*tkt_-2	Sequence variations	0.92	0.08	Yes
03/270	*tkt_-2	Sequence variations	0.965	0.044	Yes
03/414	*tkt_-2	Sequence variations	0.888	0.634	Yes
03/427	*tkt_-2	Sequence variations	0.955	0.118	Yes
03/438	*tkt_-5	Sequence variations	0.9	0.257	No
03/439	*tkt_-2	Sequence variations	0.972	0.021	Yes
03/451	tkt_-3	Sequence variations	0.894	0.342	Yes
03/460	*tkt_-5	Sequence variations	0.902	0.235	No
03/464	*tkt_-1	Sequence variations	0.879	0.986	Yes
03/467	*tkt_-2	Sequence variations	0.885	0.575	Yes
03/471	*tkt_-2	Sequence variations	0.887	0.45	Yes
03/474	*tkt_-2	Sequence variations	0.963	0.053	Yes
03/478/GB	*tkt_-5	Sequence variations	0.961	0.06	No
BAA1177(515)	tkt_-3	Sequence variations	0.958	0.051	Yes
GBS III	tkt_-5	Match	0.97	0.006	Yes
GBS IV	*tkt_-2	Sequence variations	0.973	0.004	Yes
GBS VII	tkt_-2	Match	0.977	0.003	Yes
GBS VIII	*tkt_-2	Sequence variations	0.962	0.013	Yes
GBS X	tkt_-1	Sequence variations	0.921	0.234	Yes
H03420043	*tkt_-1	Sequence variations	0.888	0.633	Yes
H034540119	tkt_-3	Sequence variations	0.933	0.128	Yes
H035140030	*tkt_-2	Sequence variations	0.957	0.094	Yes
H040200291	*tkt_-2	Sequence variations	0.888	0.44	Yes
H040540417	*tkt_-1	Sequence variations	0.888	0.752	Yes

The results show that the majority of sequences were correctly identified by the iSEQ platform with 93.3% of *atr* loci, 100% of *pheS* loci and 90% of *tkt* loci being correctly identified with an average 94.4%. This is a comparable level of accuracy to the *N. meningitidis* MLST set (91). However, this only assessed loci where sequences almost exclusively had a high sequence variation probability. When most sequence variation probabilities were high it is possible that the scoring is incorrect. However, if the proportion of confident matches to matches with a significant proportion of sequence variation is roughly even then it was unclear if the scoring was incorrect or if the sequenced loci were not present in the reference database. This is what had occurred

with the *glcK* locus with 59 confident matches and 70 matches with a significant probability of sequence variation. Therefore to investigate the accuracy of the iSEQ system further all isolates were Sanger sequenced for the *glcK* locus. The *glcK* loci iSEQ sequences matched the Sanger sequence 97.7% of the time, 3 sequences out of 130 did not match the predicted sequence type and 4 were shown to contain a 2kb insert. Table 3.8 shows the three samples where the di-deoxy sequencing did not match the iSEQ sequencing results.

Table 3.8: The non-matching sequences from sequencing all glcK loci

Strain	Loci	Sequenom ST	Confidence	Score	P Variation	Match?
03/414	glcK	glcK-3	Sequence variations	0.906	0.931	No
03/474	glcK	*glcK-2	Sequence variations	0.933	0.222	No
NEM316	glcK	*glcK-2	Sequence variations	0.95	0.05	No

Sequence variation probabilities of incorrect iSEQ allele type assignments does not reveal any particular pattern (table 3.8), for example the variation probabilities for incorrectly matched reference sequences were not all scored as very high and range from 0.05 to 0.931 and sequences that match the Sanger sequence can have sequence variation probabilities as high as 0.996. Therefore, the sequence variation probabilities were not a reliable indication of the accuracy of an identified sequence. Consequently, the sequence variation probabilities were not able to identify loci that have been incorrectly sequenced it was necessary to identify a marker to select incorrectly assigned sequence types. Out of the ten sequences identified as being incorrect by iSEQ profiling 7 of these were in new sequence types that were not present in the pubMLST database but when the correct sequence was identified through Sanger sequencing the allele type was corrected to a previously identified allele type. Therefore, it is highly likely that allelic profiles that are not present in the MLST database contain incorrectly assigned allele types. Each allelic profile not present in the MLST database was searched using the allelic profile query on the pubMLST website, this revealed a list of sequence types that had an allelic profile close to that of the submitted sequence type and allowed the identification of allele types that are likely to be

incorrectly assigned. For example, table 3.9 shows the list of allelic profiles close to the profile given for the isolate 8190 (1,1,2,8,1,2,2) which was not found in the MLST profile database.

Table 3.9: Comparison of allelic profiles similar to the profile for isolate 8190 which was not found in the database

ST	*adhP*	*pheS*	*atr*	*glnA*	*sdhA*	*glcK*	*tkt*
NEW ST 8190	1	1	2	8	1	2	2
1	1	1	2	1	1	2	2
50	1	1	2	11	1	2	2
97	1	1	2	18	1	2	2
251	1	1	2	27	1	2	2
297	1	1	2	2	1	2	2
383	1	1	2	42	1	2	2
413	1	1	2	4	1	2	2
429	1	1	2	46	1	2	2
463	1	1	2	3	1	2	2
478	1	1	2	5	1	2	2

From this analysis only the *glnA* allele differs from the allelic profile of 10 known sequence types suggesting that the *glnA* allele may have been incorrectly assigned. The same analysis was performed for all 17 allelic profiles that were not identified as a known allele type. The alleles that differed from closely related allele types were Sanger sequenced. In total 30 alleles were Sanger sequenced (9 *adhP*, 2 *atr*, 1 *glcK*, 4 *glnA*, 2 *pheS*, 3 *sdhA* and 9 *tkt*) and of these alleles, 16 matched the iSEQ allele type assignment and 14 did not match. The 14 non-matching sequences corrected the allele types of 12 previously unidentified allele types and placed them into the correct group, a further 5 previously unidentified allele types were shown to be correct and are therefore new allele types.

In conclusion, the iSEQ system can correctly identify the alleles of the MLST scheme. The platform will correctly identify 95.5% of sequences, which is comparable to other work (91). It is possible when using the MLST set to identify sequences that are more likely to be incorrectly identified. However, this approach for identifying incorrectly assigned sequences could not be applied to any

new sequence typing methods developed since there would be no existing database for comparison. Therefore, the iSEQ method was not used in further work.

3.1.3 MLST Profiles

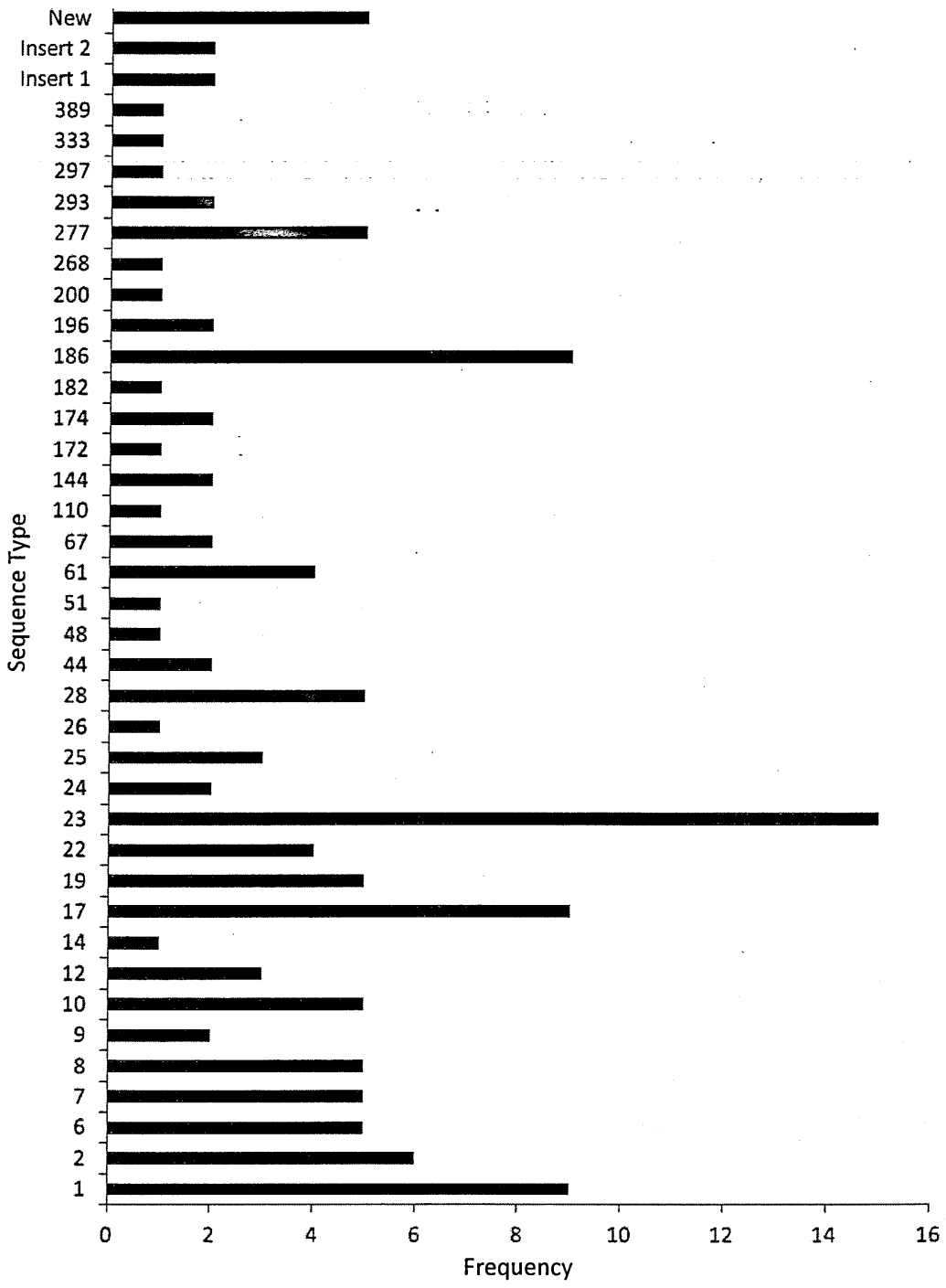
The iSEQ data combined where necessary with Sanger sequencing was used to determine the sequence type of each isolate in the strain collection using the pubMLST website (101). The 134 clinical isolates were separated into 43 unique sequence types (summary in table 3.10 and figure 3.3. Full results with allelic profiles and serotype data is in Appendix 9.2). The relationships between the sequence types is shown using a maximum likelihood tree (figure 3.4).

Table 3.10: Summary of MLST sequence types

ST	Frequency
1	9
2	6
6	5
7	5
8	5
9	2
10	5
12	3
14	1
17	9
19	5
22	4
23	15
24	2
25	3
26	1
28	5
44	2
48	1
51	1
61	4
67	2

ST	Frequency
110	1
144	2
172	1
174	2
182	1
186	9
196	2
200	1
268	1
277	5
293	2
297	1
333	1
389	1
Insert 1	2
Insert 2	2
NEW-1	1
NEW-2	1
NEW-3	1
NEW-4	1
NEW-5	1

Figure 3.3: Frequency of MLST sequence types



Unsurprisingly in a collection of mostly clinical isolates, the virulent ST-23 sequence type is the most prevalent (15/134) followed by ST-1, ST-17 and ST-186 (each with 9/134). A Chi squared statistic was calculated as 75.061 with 38 degrees of freedom which gave a P value of 0.0003, indicating a statistically significantly non-random distribution.

Figure 3.4: A maximum likelihood tree indicating the relationships between MLST sequence types using the GTR model, optimised proportion of invariable sites and optimised gamma shape parameter

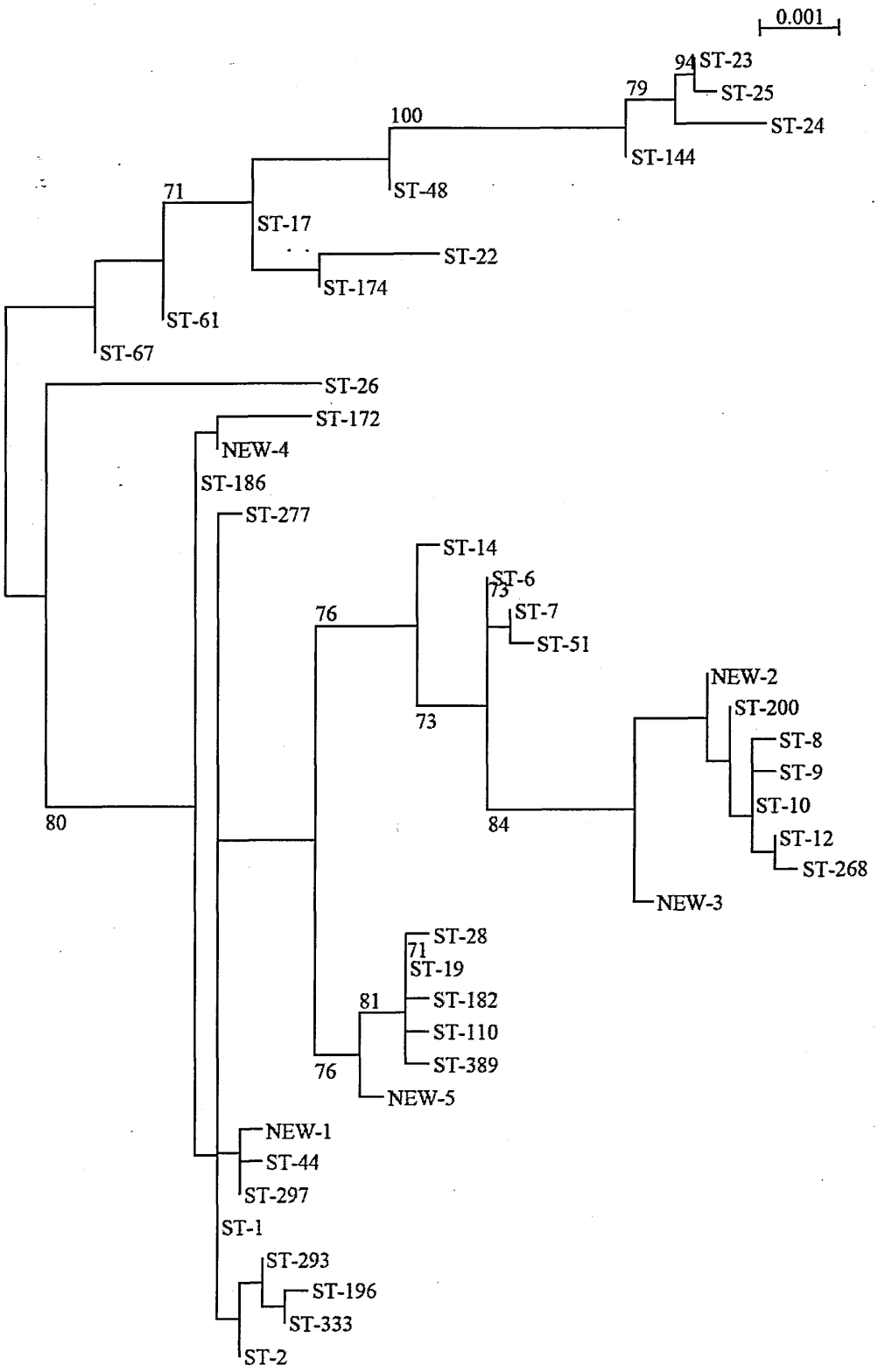


Figure 3.4 shows the relationship between the sequence types generated in this study and was calculated by performing maximum likelihood analysis on the concatenated allele sequences of all 7 loci. The 4 isolates that contained 2kb inserts in the *glcK* loci were removed from the analysis. The tree shows that the new sequence types are distributed throughout the tree. That they are not located in their own clade suggests that these 5 strains are not a novel lineage but considering the high levels of recombination observed previously between MLST loci (221) it is highly likely that there have been recombination events which created novel allelic profiles. The four most commonly isolated sequences types 1, 17, 23 and 186 are also distantly related suggesting these are separate lineages which have become prevalent independently and are not descended from a recent common ancestor. Finally, the three sequence types 24, 25 and 144 are closely related to the invasive sequence type ST-23, suggesting these may also be invasive sequence types.

3.2 Discussion of MLST Profiling

MLST is currently the gold standard molecular typing scheme for GBS (103) and is used for profiling in clinical research and epidemiology settings. According to the current version of the pubMLST database the GBS MLST scheme is reasonably well represented with 559 unique sequence types and an average of 46 unique allele types per loci (min = 45, max = 105) which places the GBS MLST scheme as the 16th largest scheme out of 80 meaning it is already well established. The largest MLST scheme is the *Neisseria* species MLST scheme which has 9060 unique sequence types and as has been shown to use alleles that are far more variable than the alleles of the GBS MLST scheme. Sequence typing of *N. menginitidis* was also the scheme used to demonstrate that the Sequenom iSEQ platform could be used for accurate sequence typing (91). The MLST scheme was used to test the iSEQ platform for sequence typing and to provide a point of reference for comparison with new sequence typing schemes developed.

The distribution of MLST sequence types generated in this study showed a higher number of unique sequence types and a lower number of isolates were placed into the top four sequence types compared to other studies (103,104). Here 31% of isolates were placed into the top 4 sequence types and 43 sequence types were identified. This suggests that the strain collection used was more variable than strain collections used in previous studies. This is understandable since previous studies used strains from single regions or hospitals, compared to this study in which the isolates came from across the UK. The most prevalent sequence type in this strain collection was the hyper-virulent ST-23 (15/134 isolates) which is unsurprising in a collection of clinical isolates and the top four sequences types in this study 1, 17, 23 and 186 are also known to be common sequence types in neonatal infection as demonstrated in other studies (17,23,103,104,125,141,152,156). Statistical analysis of the data generated in this study showed that the isolates were not randomly distributed and that the MLST data is therefore indicating some kind of evolutionary relationship. Additionally, phylogenetic analysis revealed that the four most commonly isolated sequence types are distantly related which suggests that these are separate lineages which have become prevalent independently. Potentially they have each

acquired a novel feature that increased virulence and allowed these isolates to successfully disseminate by to clonal expansion. However, since it was shown that these four sequence types are not closely related according to the MLST data that would mean that either a common feature that causes enhanced virulence was acquired by a distant common ancestor which has been conserved by strong selective pressure whilst the rest of the organism changes. It is also possible that each sequence type has independently acquired separate features that enhance virulence and allowed each sequence type to become prevalent. The final possibility is that the MLST scheme is not correctly establishing the relationships between a particularly subsets of GBS clinical isolates and hence why a new profiling method for GBS is required.

To test the potential of a new profiling method requires comparison to an existing method, in this case to MLST. Therefore the strain collection was typed using the existing MLST scheme and this was performed using the iSEQ system for comparative sequence analysis.

As shown in the results section (3.1.2) the platform had difficulty in assessing the probability of a given sequence being correct. However, Sanger sequencing confirmed the majority of the sequences generated by the iSEQ platform and showed that no new allele types were generated. This is unsurprising because although these isolates have not been previously characterised by MLST a large number of UK clinical isolates have (103,104) and these isolates would not be expected to be so different. There are several potential reasons why the iSEQ system was not able to analyse the data. Firstly it has previously been shown that AT rich regions do not sequence well in the iSEQ platform since T bases do not accept charge and fly well in a spectrometer (C. Honsich Personal Correspondence). This would make GBS a particularly poor organism to be sequenced using the iSEQ platform as the average AT content of GBS is approximately 65%, and analysis of MLST targets showed that they are on average around 10% more AT rich than those of the *N. meningitidis* MLST scheme. The second reason is that the GBS MLST scheme is far less variable than the *N. meningitidis* MLST scheme which, combined with the problems in AT rich fragments not flying well in a spectrometer means the platform will be less able to differentiate between

highly similar alleles. If there was limited variation and that variation occured in an AT rich region it would be less likely to be identified.

iSEQ sequencing did show that the vast majority of isolates were identified with comparable results to those previously generated using the Sequenom (91). Despite it being difficult to identify which isolates were sequenced correctly using the iSEQ platforms statistics it was observed that sequences that were identified incorrectly had a much higher chance of being in allele types that have not previously been identified in the online MLST allele database (101). For this reason, any isolate that was identified as a new allele type was sequenced using Sanger sequencing. From this the number of isolates found to be new allele types was reduced from 30 to 5 which is in line with previous research (91). This means that analysis of the sequencing results could correct the problems with the analysis of GBS sequences to give a similar accuracy rate to the *N. meningitidis* MLST scheme used by Honisch et al. (91). For these reasons it is believed that the MLST sequence data generated was correct and could be used as a comparison for future new targets. However, since these problems were identified in targets that already have a large reference database which allowed identification of atypical isolates, using this platform on completely novel targets developed in the next section would mean that to be sure the correct sequence type was assigned to each isolate they would have to be identified by Sanger sequencing anyway, therefore making the iSEQ platform unsuitable for further use in this project. Additionally, even towards the end of this project the costs of Sanger sequencing had decreased below the point of the initially cheaper iSEQ platform and advances in automation and capacity of Sanger sequencing technology had increased to make the Sequenom platform obsolete for sequencing.

Chapter 4

Analysis of the Core Genome

4.0 Analysis of the Core Genome

4.1 Introduction

The aim of this work was to discover phylogenetically representative target markers that better represent the relationships shared by the whole genomes of GBS. To do this a method that has previously been developed for sequence typing at the genus level (119) was adapted to develop a sequence typing method at the strain level. Reciprocal BLAST was used to extrapolate the core genomes of 1) three fully sequenced GBS genomes (2603V/R, A909 and NEM316) and 2) the three fully sequenced genomes plus five whole genome shotgun (wgs) sequenced genomes (18RS21, 515, CJB111, COH1 and H36B). The core genome was then used to generate the ANI which was correlated to distance values between each ortholog of each core gene. The aim was to create a strain typing system that was not only more representative of the whole genome but which is also able to provide comparable or improved discriminatory power when compared to the current gold standard MLST scheme (103).

4.1.1 Data Mining of Sequenced Genomes

The coding sequences from the three fully sequenced and five wgs genomes were identified (Table 4.1). The average number of genes across the 8 genomes was 2198 with a minimum of 1999 coding genes (A909) to a maximum of 2376 (COH1 and H36B).

Table 4.1: The total number of coding DNA sequences for each genome

Genome	Number of Coding Sequences	Method of Sequencing
2603V/R	2121	Fully Sequenced
A909	1996	Fully Sequenced
NEM316	2094	Fully Sequenced
515	2275	Whole Genome Shotgun
18RS21	2146	Whole Genome Shotgun
CJB111	2197	Whole Genome Shotgun
COH1	2376	Whole Genome Shotgun
H36B	2376	Whole Genome Shotgun

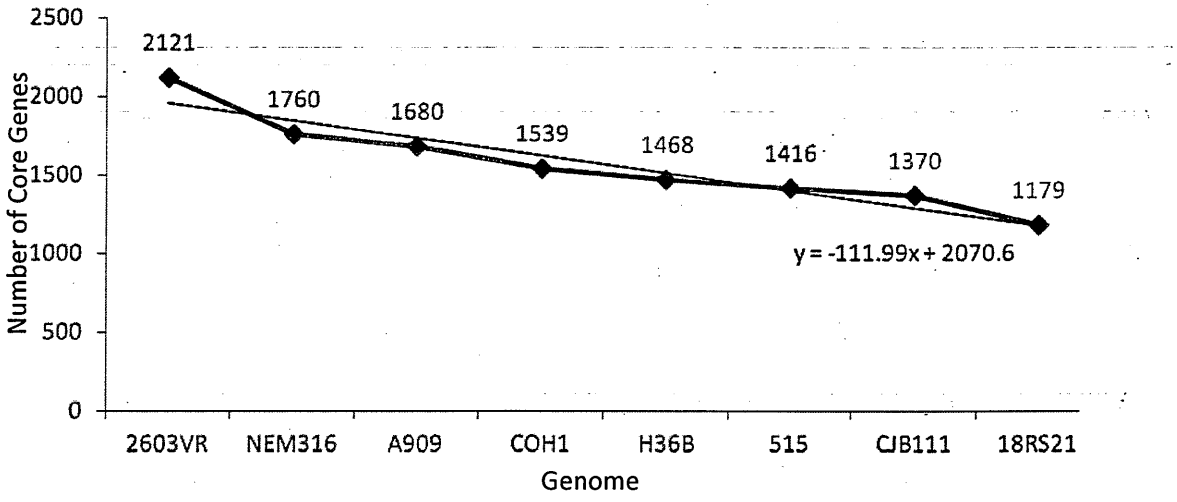
4.1.2 Reciprocal BLAST Results

Reciprocal best match BLAST is an implementation of the BLAST DNA alignment algorithm which is used to identify user generated DNA sequences by performing short alignments over a given sequence database and identifying sequences with a specified level of consecutive alignments. Reciprocal best match BLASTs a reference DNA sequence against a set of query genomes using defined length and homology criteria to determine any DNA sequences in the reference DNA sequences that meet the set length and homology criteria. Of the hits that meet the criteria, the most similar was taken as the best match. This best match was then realigned and searched against the initial reference sequence to confirm it meets the length and homology criteria. Reciprocal best match BLAST was used here to identify the core genomes of the three fully sequenced genomes and of all eight GBS genomes.

To ensure that no one genome was having a disproportionate effect on the size of the core genome (since wgs sequences have gaps and there is no guarantee that coding sequences do not occur inside those gaps) the core genome was calculated for all strains by adding in each strain sequentially. That is, 2603V/R was used as the reference sequence as it was the largest genome of the fully sequenced genomes and reciprocal BLAST was carried out on the genome NEM316. The number of genes returned as core between these two genomes was recorded. Then the core genome of all three sequenced genomes was calculated by performing reciprocal BLAST using 2603V/R as the reference and the genomes A909 and NEM316 as query genomes. Again, the number of core genes was recorded. Following this, 5 more reciprocal BLAST searches were preformed adding each of the whole genome shotgun genomes sequentially from smallest to largest (i.e., COH1, H36B, 515, CJB111, 18RS21). At each reciprocal BLAST search the number of core genes identified was recorded to confirm that no genomes from the whole genome shotgun set were missing a disproportionate number of coding sequences that would affect the size of the core genome. This showed a broadly linear trend (Figure 4.1) suggesting that the reduction in the number of core genes when new sequences are added is not a result of missing coding sequences

from the wgs genomes, although a limited number of missing coding sequences cannot be ruled out.

Figure 4.1: A graph indicating the number of core genes found when new genomes are added to calculating the core genome.



After it was shown that no one genome has a disproportionate effect on the size of the core genomes, two datasets were used for further analysis. The first dataset comprised the core genome of the three fully sequenced genomes (2603V/R, NEM316 and A909) and had a core genome containing 1684 genes and the second dataset comprised core genes from all the sequenced genomes and had a core genome containing 1179 genes (Appendix 9.10.1 and 9.10.2 for the full list of core genes for the 3 and 8 core genome datasets respectively). The proportion of core genes in each dataset was calculated as shown in tables 4.2 and 4.3. For both datasets, the largest genomes (2603V/R and COH1 and H36B) had the lowest percentage of core genes and the smallest genome (A909) has the highest level. This suggests that the core genome is quite small with a large accessory genome and as more GBS genomes are sequenced the size of the core genome will continue to fall. Theoretically, after 18.49 genomes the core genome will contain no genes, although it is obvious that at some point before that the number of core genes reach a minimum level which would cover the "true" core genome for GBS. Compared to *E. Coli* this is a rather low number of genomes before reaching the theoretical zero point. Lukjancenko et al.

(142) showed that in *E. coli* at the same number of genomes the core-genome still accounts for a large proportion of individual strains.

Table 4.2: Core genes in the fully sequenced genomes

Strain	Total Gene Content:	% Core	% Variable
2603V/R	2124	79.3	20.7
NEM316	2093	80.5	19.5
A909	1996	84.4	15.6

Table 4.3: Core genes in all sequenced genomes

Strain	Total Gene Content:	% Core	% Variable
COH1	2376	49.6	50.4
H36B	2376	49.6	50.4
515	2275	51.8	48.2
CJB111	2197	53.7	46.3
18RS21	2146	54.9	45.1
2603V/R	2124	55.5	44.5
NEM316	2093	56.3	43.7
A909	1996	59.1	40.9

Each gene in the core genome was assigned to a cluster of orthologs groups (COG) category and plotted as a percentage of the total number of genes in that COG category in the reference genome 2603V/R to identify how different COG categories are represented in the core genome (Figure 4.2 and Table 4.4).

Figure 4.2: The percentage of each COG category present in each of the two core genome datasets

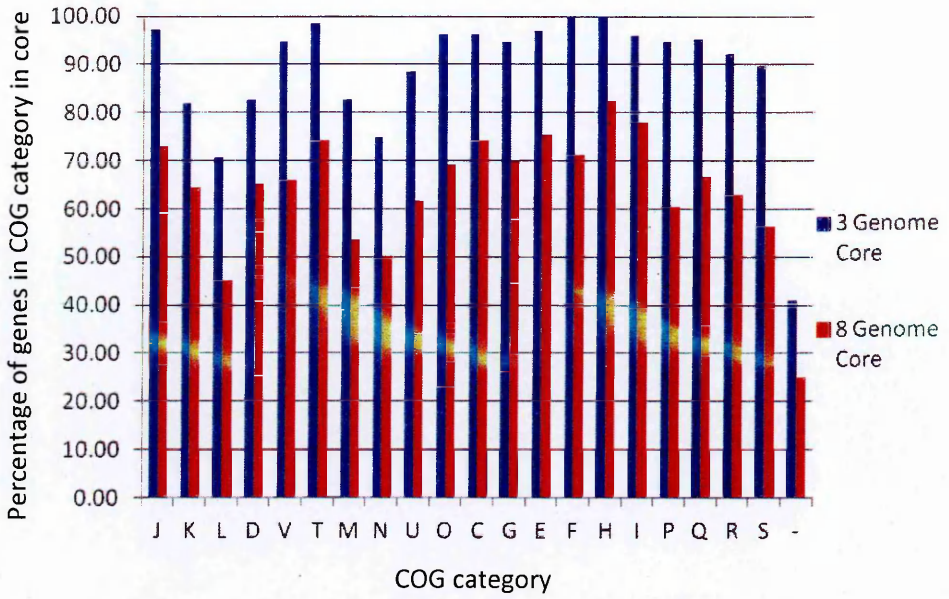


Table 4.4: The total number of genes from each category in the 2603V/R genome and the description of each category

COG	Genes in COG (2603V/R)	Description
J	152	Translation
K	160	Transcription
L	144	Replication, recombination and repair
D	24	Cell cycle control, mitosis and meiosis
V	45	Defence mechanisms
T	74	Signal transduction mechanisms
M	114	Cell wall/membrane biogenesis
N	8	Cell motility
U	27	Intracellular trafficking and secretion
O	57	Posttranslational modification, protein turnover, chaperones
C	61	Energy production and conversion
G	153	Carbohydrate transport and metabolism
E	160	Amino acid transport and metabolism
F	74	Nucleotide transport and metabolism
H	51	Coenzyme transport and metabolism
I	51	Lipid transport and metabolism
P	109	Inorganic ion transport and metabolism
Q	27	Secondary metabolites biosynthesis, transport and catabolism
R	247	General function prediction only
S	166	Function unknown
-	444	Not in COGs

The proportion of genes from each COG category was relatively evenly distributed across both the datasets with some notable exceptions. Genes that do not feature in the COG database were underrepresented in both core genomes with only 41% of genes not in the COG database present in the 3 genome core dataset and 25% in the 8 genome core dataset. In the 3 genome core dataset two COG categories have all their genes in the core (Nucleotide transport and metabolism and Coenzyme transport and metabolism) and a further 11 categories had over 90% of their genes in the core genome (Translation, Defence mechanisms, Signal transduction mechanisms, Posttranslational modification, protein turnover and chaperones, Energy production and conversion, Carbohydrate transport and metabolism, Amino acid transport and metabolism, Lipid transport and metabolism, Inorganic ion transport and metabolism, Secondary metabolites biosynthesis, transport and catabolism and General function prediction only). The 8 genome core dataset showed a relatively consistent fall in the proportion of each COG category (average 25%) with a few notable exceptions. Only two categories showed more than a 30% reduction in the proportion of COG categories found in the 8 genome core dataset opposed to the 3 genome core dataset, function unknown (33.5%) and inorganic ion transport and metabolism (34.4%) and only 5 categories showed less than a 20% reduction, transcription (17.4%), cell cycle control, mitosis and meiosis (17.4%), coenzyme transport and metabolism (17.7%), lipid transport and metabolism (18%) and not in COG (16%).

4.1.3 Alternate Core Genomes

To confirm the core genome identified here using reciprocal BLAST two separate methods were used. Firstly, the Panseq online tool (124) was used to create an alignment of the core genomes of the three fully sequenced genomes (unfortunately this method could not be applied to the five whole genome shotgun sequences). One ortholog of the alignment was then used in glimmer to identify potential open reading frames, the number of which was taken to be the core genome. This gave a core genome of 1809 sequences. However, not all predicted ORFs are necessarily actual genes. Secondly, the Multi-Genome Homology Comparison tool (JCVI) was used to identify

the core genome of both the three genome dataset and the eight genome dataset using a 50% sequence identity cut off as in the reciprocal BLAST approach. This identified a core genome of 1713 genes for the three genome dataset and a core genome of 1485 genes for the eight genome dataset.

4.1.4 ANI Calculation Results

ANI is the Average Nucleotide Identity between each ortholog in a concatenated alignment of a core genome. It has been used previously to investigate the species definition within a genus (120) and as a way to select genes that are evolving at the same rate as the core genome in development of genus level sequence typing methods (119). Here, the ANI will be correlated to the nucleotide identities of every core gene in both core genome datasets to allow selection of profiling markers at the strain level.

The level of nucleotide identity between each genome pair from both datasets was calculated and is shown in the distance tables 4.5 and 4.6 respectively.

Table 4.5: The nucleotide identities of the core genome pairs of the three genome core dataset

		1	2	3
1	2603V/R	100		
2	NEM316	98.1	100	
3	A909	97.9	98.4	100

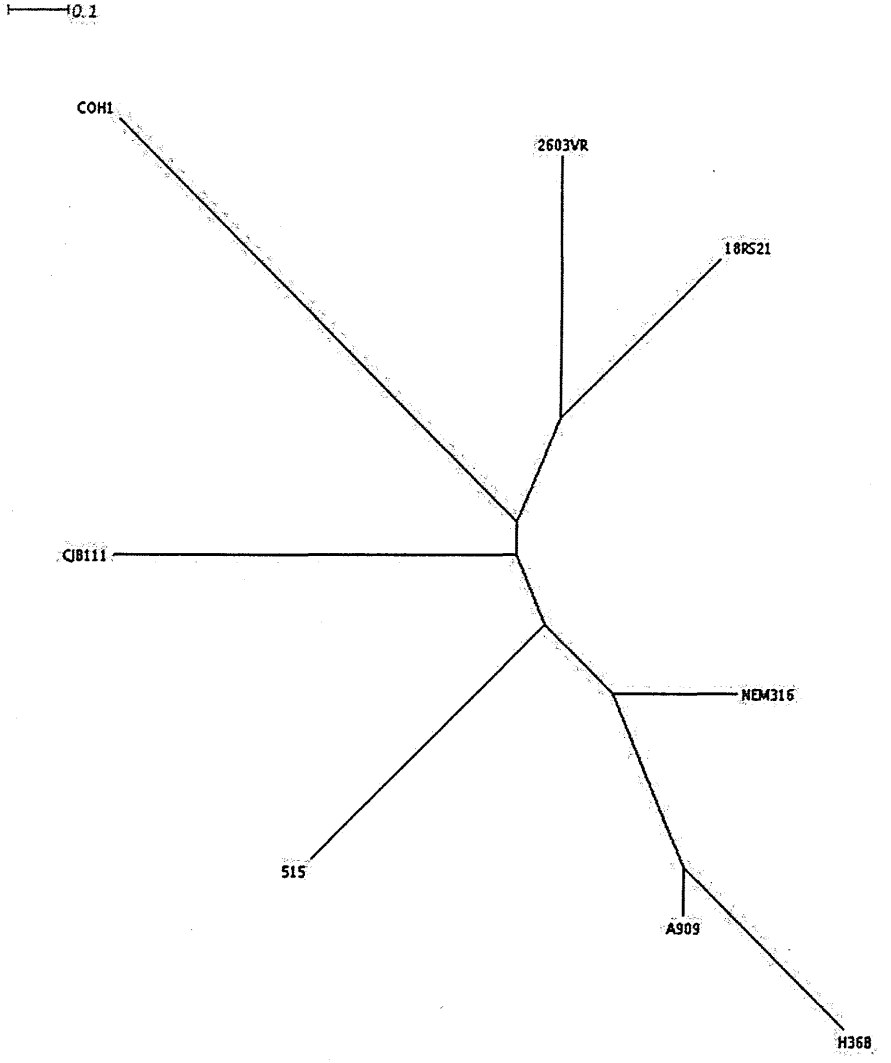
Table 4.6: The nucleotide identities of the 8 genome core dataset

		1	2	3	4	5	6	7	8
1	515	100							
2	18RS21	98.77	100						
3	2603VR	98.66	99.19	100					
4	A909	98.48	98.64	98.56	100				
5	CJB111	98.57	98.7	98.64	98.99	100			
6	COH1	98.34	98.43	98.49	98.35	98.33	100		
7	H36B	98.30	98.61	98.36	99.54	98.85	98.21	100	
8	NEM316	99.29	98.92	98.78	99.54	98.68	98.42	98.42	100

From these values the ANI was calculated by averaging the distance values between each genome pair. For the three genome core dataset this was 98.11% and for the 8 genome core dataset was 98.68%. An ANI value of >94% corresponds to the traditional species definition of 70% DNA-DNA re-association and genomes from the same species have been shown to have ANI values between 98 and 99% (120). This is confirmed for the GBS genomes.

The ANI values of the core genome of all 8 strains were converted into a Phylip formatted DNA distance matrix and used in the BioNJ program (67) to generate a neighbour joining tree of the core genomes which essentially clusters the isolates based on the hierarchy of the level of identity between isolates calculated based on the distance values (Figure 4.3).

Figure 4.3: A neighbour joining tree of the core genome of GBS.



As shown in the tree the most closely related genomes are A909 to H36B and NEM316. The core genomes of H36B and NEM316 share 99.54% homology to core genome of A909. The most distantly related genomes are H36B and COH1 which are 98.21% homologous. The genome sequenced strains represent seven GBS serotypes. Three of these serotypes appear twice in the tree (Ia, III and V). It is therefore possible to speculate on the correlation between genomic relationship and serotype for these three serotypes. The strains A909 and 515 are both serotype Ia and are separated by three common ancestors as are the serotype V strains 2603VR and CJB111. The serotype III strains COH1 and NEM316 are separated by four common ancestors. The higher levels of variation shown between genomes sharing the same serotype indicate, that serotype is not correlated to relationships of whole genomes as suggested by other studies (150,219).

4.1.5 Maximum Likelihood

Maximum likelihood analysis was performed for each core gene from both datasets as previously done by Konstantinidis et al. (119) using a custom Perl script (appendix 9.8.5) to automate the use of PAUP* and ModelTest v3.7. The evolutionary distance values from this process were used to correlate the relationship between each core gene ortholog to the ANI to aid in the selection of phylogenetically representative targets.

Of the 1684 genes from the 3 genome core dataset 1341 gave evolutionary distance values. From the 8 genome dataset 1119/1179 core genes gave evolutionary distance values. The remaining core genes (343 for the 3 genome dataset and 60 for the 8 genome dataset) could not be analysed due to high levels of homology between each ortholog and where either identical across all loci or differed by only 1bp. These loci were therefore removed from further analysis since loci that display a high level of homology would make poor profiling markers.

4.1.6 Kendal's τ Rank Correlation Co-efficient

The Kendal's τ rank correlation co-efficient is a non-parametric hypothesis test and is used to measure the association between two quantities. Specifically, it is a measure ranking correlation, i.e. the correlation between the rank given to a piece of data in two separate lists. Here, the distance value between each pair of orthologs of the ANI is ordered and is correlated to the rank order of each core gene. Genes that are highly correlated to the ANI using the Kendal's τ rank correlation co-efficient are therefore evolving in the same way as the average for the core genome and should provide an accurate measure of phylogeny.

The distance values from each gene in both the three genome and eight genome core datasets were correlated to the ANI using the Kendal's τ rank correlation co-efficient. There are two Kendal's τ rank correlation co-efficient scores, tau-a and tau-b, the stata software package calculates both. However, the tab-b score was used for target selection since it corrects for ties between values and is suited to square tables, i.e. distance matrix. Tau-a/b scores range from 1 (100% correlation) to -1 (0% correlation).

Correlation of the 3 genome core dataset to its ANI split 1341 core genes into 7 separate groups (the maximum mathematically possible considering the number of different potential ranks compared to the ANI). From this 66 core genes were placed into the top scoring tau-1 group (Table 4.7).

Table 4.7: All core genes from the 3 genome core dataset that have a Kendal's τ score of 1

Locus Tag	Product Name	Locus
SAG1305	homocysteine methyltransferase	*mmuM*
SAG0445	valyl-tRNA synthetase	*valS*
SAG1241	IS3 family transposase OrfA	-
SAG2110	50S ribosomal protein L33	*rpmG*
SAG1334	peptide deformylase	-
SAG1145	sodium:alanine symporter family protein	-
SAG0705	glycosy hydrolase family protein	-
SAG1777	ribosome-associated GTPase	-
SAG1085	xanthine permease	*pbuX*
SAG1070	ABC transporter, ATP-binding/permease protein	-
SAG0420	ribonucleotide-diphosphate reductase subunit alpha	*nrdE-1*
SAG1320	hypothetical protein SAG1320	-
SAG1193	TPR domain-containing protein	-
SAG0685	hypothetical protein SAG0685	-
SAG1372	thiamine biosynthesis protein ThiI	*thiI*
SAG0137	hypothetical protein SAG0137	-
SAG0142	hypothetical protein SAG0142	-
SAG0047	adenylosuccinate lyase	*purB*
SAG1321	hypothetical protein SAG1321	-
SAG0096	heat shock protein GrpE	*grpE*
SAG1307	hypothetical protein SAG1307	-
SAG0701	glucuronate isomerase	*uxaC*
SAG0107	CTP synthetase	*pyrG*
SAG1096	hypothetical protein SAG1096	-
SAG1160	UDP-N-acetylglucosamine-2-epimerase NeuC	*neuC*
SAG1785	hypothetical protein SAG1785	-
SAG1171	glycosyl transferase CpsE	*cpsE*
SAG0696	sugar transporter, putative	-
SAG1172	cpsD protein	*cpsD*
SAG1403	hypothetical protein SAG1403	-
SAG1047	orotidine 5'-phosphate decarboxylase	*pyrF*
SAG2106	hypothetical protein SAG2106	-
SAG1108	spermidine/putrescine ABC transporter, spermidine/putrescine-binding protein	*potD*
SAG1319	hemolysin III, putative	-
SAG1736	x-prolyl-dipeptidyl aminopeptidase	*pepX*
SAG1077	peptide chain release factor 1	*prfA*
SAG1094	inorganic polyphosphate/ATP-NAD kinase	*ppnK*
SAG1117	folylpolyglutamate synthase	*folC*
SAG1323	isopentenyl pyrophosphate isomerase	-
SAG0405	protein of unknown function/lipoprotein, putative	-
SAG1154	DNA topoisomerase IV subunit B	*parE*
SAG0102	hypothetical protein SAG0102	-
SAG1155	putative glycerol-3-phosphate acyltransferase PlsY	-
SAG1118	rarD protein	*rarD*
SAG0406	UTP-glucose-1-phosphate uridylyltransferase	*galU*
SAG1137	gls24 protein, putative	-
SAG1202	hypothetical protein SAG1202	-
SAG0123	DNA-binding response regulator	-
SAG0135	amino acid ABC transporter, ATP-binding protein	-
SAG1115	dihydropteroate synthase	*folP*
SAG1246	hypothetical protein SAG1246	-
SAG0392	cell wall surface anchor family protein	-
SAG1198	dTDP-glucose 4,6-dehydratase	*rfbB*
SAG0228	hypothetical protein SAG0228	-
SAG0043	phosphoribosylamine--glycine ligase	*purD*
SAG0025	phosphoribosylformylglycinamidine synthase, putative	-
SAG0027	phosphoribosylaminoimidazole synthetase	*purM*
SAG0086	putative lipoprotein	-
SAG0566	prophage LambdaSa1, single-strand binding protein	*ssb-2*
SAG0606	hypothetical protein SAG0606	-
SAG1984	hypothetical protein SAG1984	-
SAG1863	prophage LambdaSa2, single-strand binding protein	*ssb-4*
SAG1980	ABC transporter, ATP-binding protein	-
SAG1982	Cro/CI family transcriptional regulator	-
SAG1245	hypothetical protein SAG1245	-
SAG1971	hypothetical protein SAG1971	-

Performing the Kendal's τ test to correlate the 8 genome core dataset to the ANI provided better discrimination splitting 1119 core genes into 826 unique tau-b scores however, no tau-b scores were perfectly correlated to the ANI. Table 4.8 shows the top 15 profiling markers.

Table 4.8: The top 15 potential profiling markers from the 8 genome core dataset

Locus Tag	Product Name	Locus	tau-a score	tau-b score
SAG1470	GTPase ObgE	*obgE*	0.6587301	0.7080954
SAG1894	cyclic nucleotide-binding domain-containing protein	-	0.5820106	0.69342
SAG2058	major facilitator family protein	-	0.5820106	0.69342
SAG1816	hypothetical protein SAG1816	-	0.5714286	0.6885933
SAG1808	LacI family sugar-binding transcriptional regulator	-	0.5555556	0.677296
SAG1834	alkyl hydroperoxide reductase, subunit F	*ahpF*	0.5079365	0.6682031
SAG1970	hypothetical protein SAG1970	-	0.5079365	0.6682031
SAG1824	Hsp33-like chaperonin	*hslO*	0.5899471	0.6565586
SAG1033	FtsK/SpoIIIE family protein	-	0.6481481	0.655986
SAG2033	acetyltransferase	-	0.5079365	0.6478174
SAG1392	iron compound ABC transporter, ATP-binding protein	-	0.5767196	0.6407914
SAG1965	phosphate ABC transporter, permease protein	-	0.5767196	0.632602
SAG1825	NifR3/Smm1 family protein	-	0.5291005	0.6303818
SAG2047	hypothetical protein SAG2047	-	0.5291005	0.6303818
SAG1826	deoxynucleoside kinase family protein	-	0.5793651	0.6285582

4.1.7 Absolute Subtraction

Since the three genome core dataset gave 66 top scoring core genes a further method was required to provide further discrimination. In this method the distance values for the ANI and each core gene were totalled and the total distance value for each core gene was absolutely subtracted from the total distance value of the ANI. This means that absolute subtraction (Abs) scores close to zero are closer to the ANI and are therefore evolving at a rate closer to the whole genome average. Abs scores were calculated using a custom Perl script (appendix 9.8.8) for all core gene distance values for both datasets. The Kendal's Tau scores were still the primary method of target selection, therefore for the three genome dataset, potential profiling markers were selected from the 66 genes with a Kendal's Tau score of 1 and an Abs score close to zero. The Abs scores from the eight genome dataset were used to confirm the Kendal's Tau scores. Tables 4.9 and 4.10 shown the top 15 targets selected by absolute subtraction and the Kendal's τ test for the three genome and eight genome dataset respectively.

Table 4.9: The Kendal's τ score and absolute subtraction score of the top 15 targets from the three genome core dataset

Locus Tag	Product Name	Locus	tau-a score	tau-b score	Abs Score
SAG1305	homocysteine methyltransferase	*mmuM*	1	1	0.0114
SAG0445	valyl-tRNA synthetase	*valS*	1	1	0.0120
SAG1241	IS3 family transposase OrfA	-	1	1	0.0126
SAG2110	50S ribosomal protein L33	*rpmG*	1	1	0.0158
SAG1334	peptide deformylase	-	1	1	0.0174
SAG1145	sodium:alanine symporter family protein	-	1	1	0.0224
SAG0705	glycosy hydrolase family protein	-	1	1	0.0258
SAG1777	ribosome-associated GTPase	-	1	1	0.0259
SAG1085	xanthine permease	*pbuX*	1	1	0.0298
SAG1070	ABC transporter, ATP-binding/permease protein	-	1	1	0.0300
SAG0420	ribonucleotide-diphosphate reductase subunit alpha	*nrdE-1*	1	1	0.0302
SAG1320	hypothetical protein SAG1320	-	1	1	0.0303
SAG1193	TPR domain-containing protein	-	1	1	0.0304
SAG0685	hypothetical protein SAG0685	-	1	1	0.0312
SAG1372	thiamine biosynthesis protein ThiI	*thiI*	1	1	0.0318

Table 4.10: The Kendal's τ score and absolute subtraction score of the top 15 targets from the eight genome core dataset

Locus Tag	Product Name	Locus	tau-a score	tau-b score	Abs Score
SAG1470	GTPase ObgE	*obgE*	0.6587301	0.7080954	0.1258582
SAG1894	cyclic nucleotide-binding domain-containing protein	-	0.5820106	0.69342	0.1713396
SAG2058	major facilitator family protein	-	0.5820106	0.69342	0.1856244
SAG1816	hypothetical protein SAG1816	-	0.5714286	0.6885933	0.0731777
SAG1808	LacI family sugar-binding transcriptional regulator	-	0.5555556	0.677296	0.1763721
SAG1834	alkyl hydroperoxide reductase, subunit F	*ahpF*	0.5079365	0.6682031	0.1215686
SAG1970	hypothetical protein SAG1970	-	0.5079365	0.6682031	0.1960784
SAG1824	Hsp33-like chaperonin	*hslO*	0.5899471	0.6565586	0.0687286
SAG1033	FtsK/SpoIIIE family protein	-	0.6481481	0.655986	1.7331175
SAG2033	acetyltransferase	-	0.5079365	0.6478174	0.1316873
SAG1392	iron compound ABC transporter, ATP-binding protein	-	0.5767196	0.6407914	0.1060607
SAG1965	phosphate ABC transporter, permease protein	-	0.5767196	0.632602	0.143535
SAG1825	NifR3/Smm1 family protein	-	0.5291005	0.6303818	0.1005128
SAG2047	hypothetical protein SAG2047	-	0.5291005	0.6303818	0.1163121
SAG1826	deoxynucleoside kinase family protein	-	0.5793651	0.6285582	0.1064163

Interestingly, the top targets by the Kendal's τ score of the 8 genome dataset had significantly higher Abs scores than the top scoring targets of the three genome dataset suggesting that the highest scoring genes from the eight genome dataset were evolving at a faster or slower rate than

the average. However, this is more likely because the top scoring group for the three genome dataset contains 70 genes and therefore a wider range of distance values are available.

4.1.8 Bioinformatic Target Selection

The top two targets based on the Kendal's Tau scores and the Absolute Subtraction scores from each core genome dataset, *valS* and *mmuM* from the three genome dataset and *obgE* and SAG1894 from the eight genome dataset and two core virulence genes *cylB* and *cpsL* from a previous study (261) were selected for sequence analysis. The levels of nucleotide variation per 100 base pairs and the number of unique clusters for the orthologs of each core gene was calculated and compared to the nucleotide variation and the number of unique clusters for the core gene orthologs of the MLST genes (Table 4.11).

Table 4.11: The bioinformatically selected targets compared to the MLST loci.

Gene	Dataset	SNPs/100bp	Unique Allele Types
mmuM	3 Genome Core	1.89	5
valS	3 Genome Core	0.79	7
obgE	8 Genome Core	0.72	5
SAG1894	8 Genome Core	0.61	3
cylB	Virulence Loci	0.12	4
cpsL	Virulence Loci	1.61	6
adhP	MLST	0.73	6
atr	MLST	0.85	5
glcK	MLST	0.35	2
glnA	MLST	0.37	3
pheS	MLST	0.34	2
sdhA	MLST	0.88	3
tkt	MLST	0.59	4

This shows that the 4 targets selected through the bioinformatic approach are more variable than all but two of seven MLST loci and the targets selected from the three genome dataset are more variable than the targets selected using the eight genome dataset. However, there is not necessarily a correlation between nucleotide variation and the number of unique allele types, for

example, the target with the highest level of nucleotide variation per 100bp (*mmuM*) will split into 5 unique allele types whereas *valS* which has nucleotide variation per 100bp, which is less than half that of *mmuM*, splits the eight genome sequenced orthologs into 7 unique allele types. This is based on a limited amount of available sequence data and by sequencing the targets from a larger strain collection the benefit of applying a bioinformatic approach to selecting sequence typing markers will be informed.

In conclusion, this section has shown that the methods used to establish the core genome and analyse each gene compared to the core genome can identify targets more variable in terms of SNPs and in terms of the number of unique allele types that the loci of the MLST scheme. It has also revealed information about the gene contents of the core genome as additional sequences are added and potentially highlighted an issue with using incomplete genome sequences when identifying the core genome. However, a bioinformatics approach alone, using a relatively small number of genomes is not sufficient to design a new sequence-typing system and the potential targets identified here need to be tested in a collection of clinical isolates.

In conclusion, this section has shown that the methods used to establish the core genome and analyse each gene compared to the core genome can identify targets more variable in terms of SNPs and in terms of the number of unique allele types that the loci of the MLST scheme. It has also revealed information about the gene contents of the core genome as additional sequences are added and potentially highlighted an issue with using incomplete genome sequences when identifying the core genome. However, a bioinformatics approach alone, using a relatively small number of genomes is not sufficient to design a new sequence-typing system and the potential targets identified here need to be tested in a collection of clinical isolates.

4.2 Sequence Analysis of the Core Genome

As mentioned, a purely bioinformatics approach using a limited number of genomes would not alone be sufficient to develop a new profiling scheme. And since the reason for developing this new profiling scheme is to develop a tool that will better represent the relationships between clinical isolates each target was tested against a collection of clinical isolates similar to what would be expected to be analysed in a national reference laboratory.

4.2.1 Core Genome Sequencing

To test the discriminatory power of the 6 targets selected using a bioinformatics approach the targets were sequenced for 79 isolates from the collection. The level of nucleotide identity and the number of unique allele types were compared to the loci of the MLST scheme (Table 4.12). The new targets split the 79 isolates into either equivalent or more unique allele types than all MLST loci except *adhP* which resulted in 9 unique sequence types whereas *mmuM* and *obgE* formed only 8 and 7 unique allele types respectively. However, the best performing loci using these criteria was *cpsL* which was selected from a set of core genome virulence genes which split 79 isolates into 18 unique allele types. The next best performing targets were *valS* (12 unique allele types) and SAG1894 (10 unique allele types) which were from the 3 genome and 8 genome datasets respectively. The worst selected target was *cylB* which split 79 isolates into only five unique sequence types and also had the highest level of nucleotide identity out of all sequenced loci. This loci was therefore removed from further analysis.

Table 4.12: The level of nucleotide identity and the number of unique allele types between each sequenced loci

Loci	Length (bp)	Unique Allele Types	Percent Identity	Source
adhP	498	9	97.99	MLST
atr	501	6	98.00	MLST
glcK	459	6	98.04	MLST
glnA	498	7	98.59	MLST
pheS	501	4	98.80	MLST
sdhA	519	6	97.50	MLST
tkt	480	7	98.13	MLST
cpsL	502	18	85.66	Core Virulence
cylB	501	5	99.21	Core Virulence
valS	476	12	95.17	3 Genome Core
mmuM	429	8	93.94	3 Genome Core
obgE	512	7	98.05	8 Genome Core
SAG1894	451	10	97.56	8 Genome Core

The data also showed that the relationship between nucleotide identity and the number of unique allele types is not clear; that is, a higher level of nucleotide variation does not always lead to a larger number of unique allele types. For example, the target SAG1894 is the third best performing target in terms of unique allele types but only the 5th most variable target with two targets showing higher levels of nucleotide variation but splitting the isolates into significantly lower levels of unique allele types (*mmuM* and *sdhA*).

The sequences from each dataset were subsequently concatenated and clustered to identify which dataset produces the most discriminatory targets (Table 4.13). Analysis showed that the genes selected from the three genome core dataset split the 79 isolates into 20 unique allele types compared to the to the eight genome dataset which split the same isolates into 15 unique allele types. The targets selected from the three genome dataset also show a higher level of nucleotide variation than the genes selected from the eight genome core dataset. Additionally, all four targets were concatenated and clustered, which showed that these four targets had equal discriminatory power to the MLST set.

Table 4.13: The number of unique allele types and the percent identity of each concatenated set of loci

Loci	Length (bp)	Unique Allele Types	Percent Identity
MLST	3456	31	98.15
mmuM, valS, obgE, SAG1894	1868	31	96.25
mmuM, valS	905	20	94.59
obgE, SAG1894	963	15	97.82

All four selected targets concatenated together have equal discriminatory power to the MLST scheme. However, the aim of this work was to create a three gene typing system that represents the phylogeny of the whole genome but also has improved discriminatory power to the MLST scheme. This combination of loci shows equal discrimination to the MLST scheme but selection of alternative markers may improve discriminatory power further. The targets from the three genome dataset are more discriminatory than the targets selected from the eight genome dataset and the three genome core dataset has 66 genes which were placed into the top scoring Kendal's Tau test category therefore the further targets were selected from this dataset. All of these 66 genes with an alignment length greater than 500bp (n= 50) were analysed to determine the level of SNP's per 100bp. For any genes that were also in the 8 genome core dataset the orthologs where used to determine the number of SNPs and for genes in the three genome dataset only, the number of SNPs was calculated from the three fully sequenced genome orthologs (Table 4.14). The level of nucleotide variation was used as an additional method for target selection since it was expected that higher levels of variation would make more discriminatory targets.

Table 4.14: The level of nucleotide variation between orthologs for the 50 genes > 500bp from the top scoring Kendal's Tau group of the 3 genome core genome dataset.

Locus Tag	Gene Name	Number of Genomes	Alignment Length (bp)	Average SNP's	SNP/100bp
SAG0027	*purM*	8	915	53.8	5.88
SAG0043	*purD*	8	1263	49.2	3.90
SAG0025	-	8	3612	124.8	3.46
SAG0047	*purB*	8	1299	32	2.46
SAG1198	*rfbB*	8	1047	23.9	2.28
SAG1305	*mmuM*	8	945	17.9	1.89
SAG1096	-	8	591	7.3	1.24
SAG1145	-	3	1346	15.3	1.14
SAG0137	-	3	1884	20.7	1.10
SAG0420	*nrdE-1*	8	2169	19.4	0.89
SAG0685	-	3	1419	12	0.85
SAG1372	*thiI*	8	1155	9.6	0.83
SAG1777	-	8	831	6.7	0.81
SAG0445	*valS*	8	1965	15.6	0.79
SAG0705	-	8	1791	13.4	0.75
SAG1070	-	8	1734	12.5	0.72
SAG1321	-	8	855	6.1	0.71
SAG1085	*pbuX*	8	1275	9	0.71
SAG0107	*pyrG*	3	1605	11.3	0.70
SAG1193	-	8	1227	8.5	0.69
SAG1160	*neuC*	3	1155	8	0.69
SAG1320	-	8	915	6.3	0.69
SAG1171	*cpsE*	8	1107	6.8	0.61
SAG1047	*pyrF*	8	702	4.3	0.61
SAG0142	-	8	1263	7.6	0.60
SAG0696	-	3	1551	9.3	0.60
SAG0701	*uxaC*	8	1401	8.2	0.59
SAG1172	*cpsD*	8	544	2.8	0.51
SAG1307	-	8	651	3.3	0.51
SAG1154	*parE*	8	1929	9.5	0.49
SAG1202	-	8	789	3.8	0.48
SAG1319	-	8	564	2.7	0.48
SAG1403	-	8	585	2.8	0.48
SAG1117	*folC*	8	1182	5.6	0.47
SAG0405	-	3	1044	4.7	0.45
SAG1155	-	3	624	2.7	0.43
SAG2106	-	8	945	3.9	0.41
SAG1736	*pepX*	8	2286	9.1	0.40
SAG1077	*prfA*	8	1080	4.1	0.38
SAG1108	*potD*	8	1074	4	0.37
SAG1137	-	3	543	2	0.37
SAG0406	*galU*	3	900	3.3	0.37
SAG1323	-	8	996	3.6	0.36
SAG0123	-	3	672	2	0.30
SAG0392	-	8	1489	4.4	0.30
SAG1118	*rarD*	8	888	2.6	0.29
SAG1115	*folP*	8	804	2.2	0.27
SAG0135	-	3	741	2	0.27
SAG1785	-	3	1293	2.17	0.17
SAG1246	-	3	1170	1.3	0.11

The four most variable genes from this list were selected for further analysis (SAG0027, SAG0043, SAG0025 and SAG0047, highlighted in red in Table 4.14) and primers were designed to amplify the most variable regions of the gene. The previously sequenced targets *mmuM* and *valS* were also found in this list at position 6 and 14 respectively, suggesting it is unlikely that there are more variable targets than the targets selected here and the targets already analysed.

The new targets were sequenced for the same 79 isolates. The nucleotide identity and the number of unique allele types were calculated for each target individually and for all 4 targets concatenated, and compared to the MLST scheme (Table 4.15).

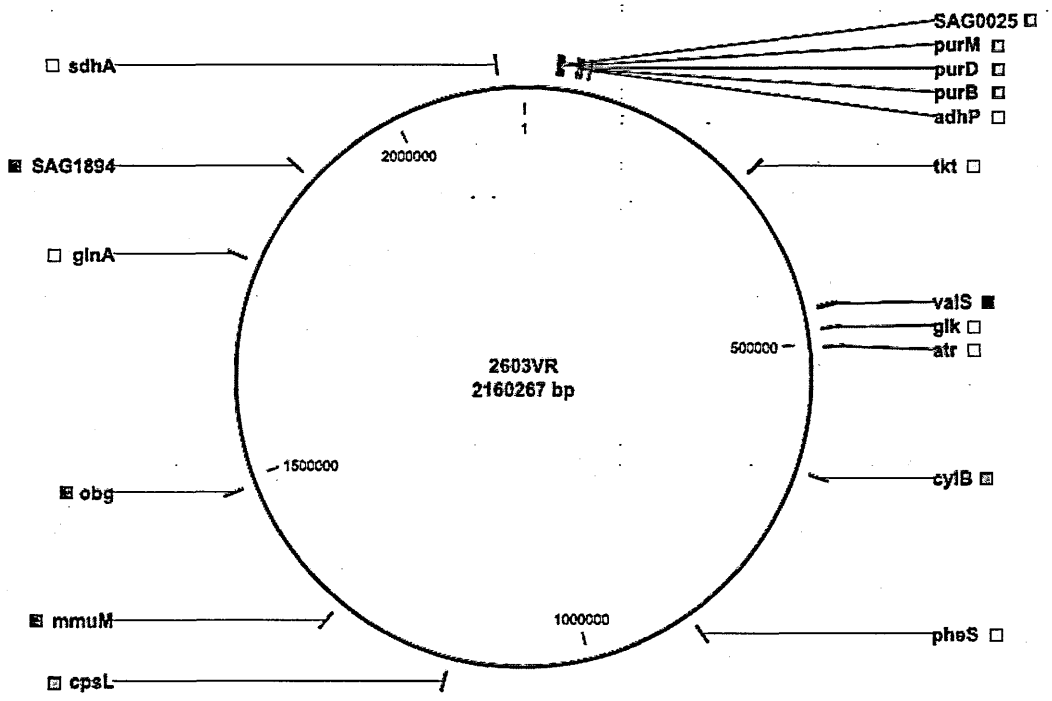
Table 4.15: Analysis of the four highly variable selected targets

Loci	Length (bp)	Unique Allele Types	Percent Identity
SAG0025	501	9	87.82
SAG0027 (*purM*)	506	8	82.81
SAG0043 (*purD*)	513	12	85.96
SAG0047 (*purB*)	488	10	91.80
Concatenated Loci	2008	22	87.05

Surprisingly, the four highly variable targets concatenated together produce less unique allele types than the MLST set (22 compared to 31 for MLST). They also produce less allele types than the concatenated *mmuM, valS, obgE* and SAG1894 targets (22 compared to 31) despite the four highly variable targets forming on average more unique allele types and having a higher level of nucleotide variation.

The locus tags suggest the four highly variable genes must be located relatively close to each other on the genome. A GenBank file modified to contain only annotations for the 17 loci sequenced in this study was imported into CLC Sequence Viewer (CLC bio) to generate a chromosome map (Figure 4.4).

Figure 4.4: A chromosome map of all sequenced targets. Red indicates the highly variable genes, Yellow indicates the MLST loci, Blue indicates the initial screening targets and Green indicates the core virulence genes.



The map shows that the four highly variable genes are located very close together on the chromosome compared to the loci of the MLST set and the other 4 core genome targets. This suggests that these four targets were evolving at the same rate due to their proximity on the chromosome and that they are all part of the purine biosynthetic pathway (188). The high levels of nucleotide variation observed do produce a higher number of allele types but when concatenated produced a limited number of very well defined sequence types.

4.2.2 Selection of 3 Profiling Marker Genes

The combination of markers used was not more discriminatory than the MLST set. However, the majority of the markers selected were more variable and form more unique clusters than the targets in the MLST set. Since at this point a large number of potential profile markers had already been sequenced at this point it was decided to combine each of the new markers to determine the optimum combination for enhanced discriminatory power using fewer genetic markers. The initial clustering data was used to select the first marker (i.e. *cpsL* formed most unique allele types and was therefore the first marker for the proposed profiling system). Every combination of *cpsL* and another one of nine of the sequenced alleles (*cylB* was excluded due to its extremely low levels of variation and poor discriminatory power) was concatenated together and clustered using the CD hit program. The sequence that generated the most new unique allele types was selected as the second profiling marker, this process was repeated until each of the 9 genes had been concatenated and clustered. This allowed the selection of the ideal number of markers by measuring the effect of the addition of a new marker to the profiling scheme (Figure 4.5). Additionally, every combination of targets was concatenated and clustered for a three and four gene typing system which showed the same markers were selected using this alternate clustering method. The sequential clustering results are shown here as they allow more discussion of the sequence typing targets and they are able to show falling returns on sequencing new targets.

Figure 4.5: The number of unique sequence types generated by concatenated profiling markers

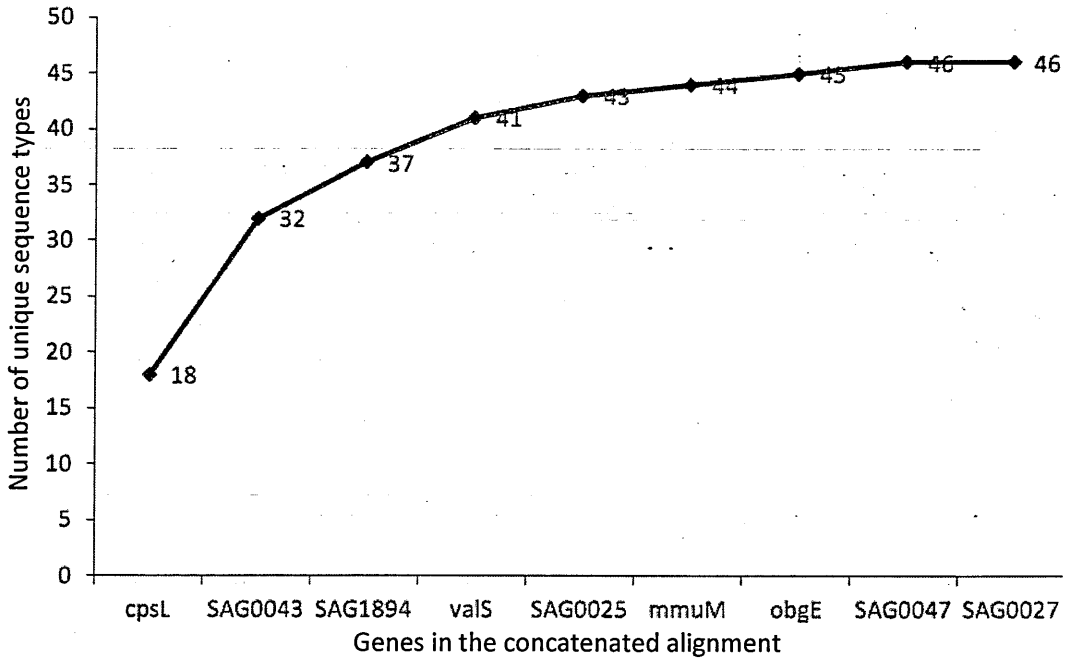


Figure 4.5 shows that the number of new sequence types starts to be reduced at the addition of SAG1894 (37 unique sequence types, 5 new) and *valS* (41 unique sequence types, 4 new) and addition of further new targets creates at best 2 new sequence types (SAG0025) and at worst no new sequence types (SAG0027).

To inform the number of profiling markers to use, the targets *cpsL,* SAG0043, SAG1894 and *valS* were sequenced for the remaining 55 isolates in the strain collection. A three gene profiling method (consisting of *cpsL,* SAG0043 and SAG1894) was compared to a four gene profiling method (three targets above plus *valS*). The four gene approach was included since it is unclear from the clustering of 79 isolates the benefit of the addition of the *valS* loci. Sequencing this fourth locus from a larger number of isolates would confirm if a three gene system was sufficiently discriminatory. These results showed that the addition of the *valS* locus split 134 isolates into 69 sequence types compared to 61 for the three gene profiling method alone (summary shown in figures 4.6 and 4.7 and the full results by isolate are shown in Appendix 9.3

and 9.4). The sequence types of the three gene profiling method that were split into new sequence types by the addition of *valS* are highlighted in green in Figure 4.6.

Figure 4.6:Summary of sequence typing using the loci cpsL, SAG1894 and SAG0043 (purD).

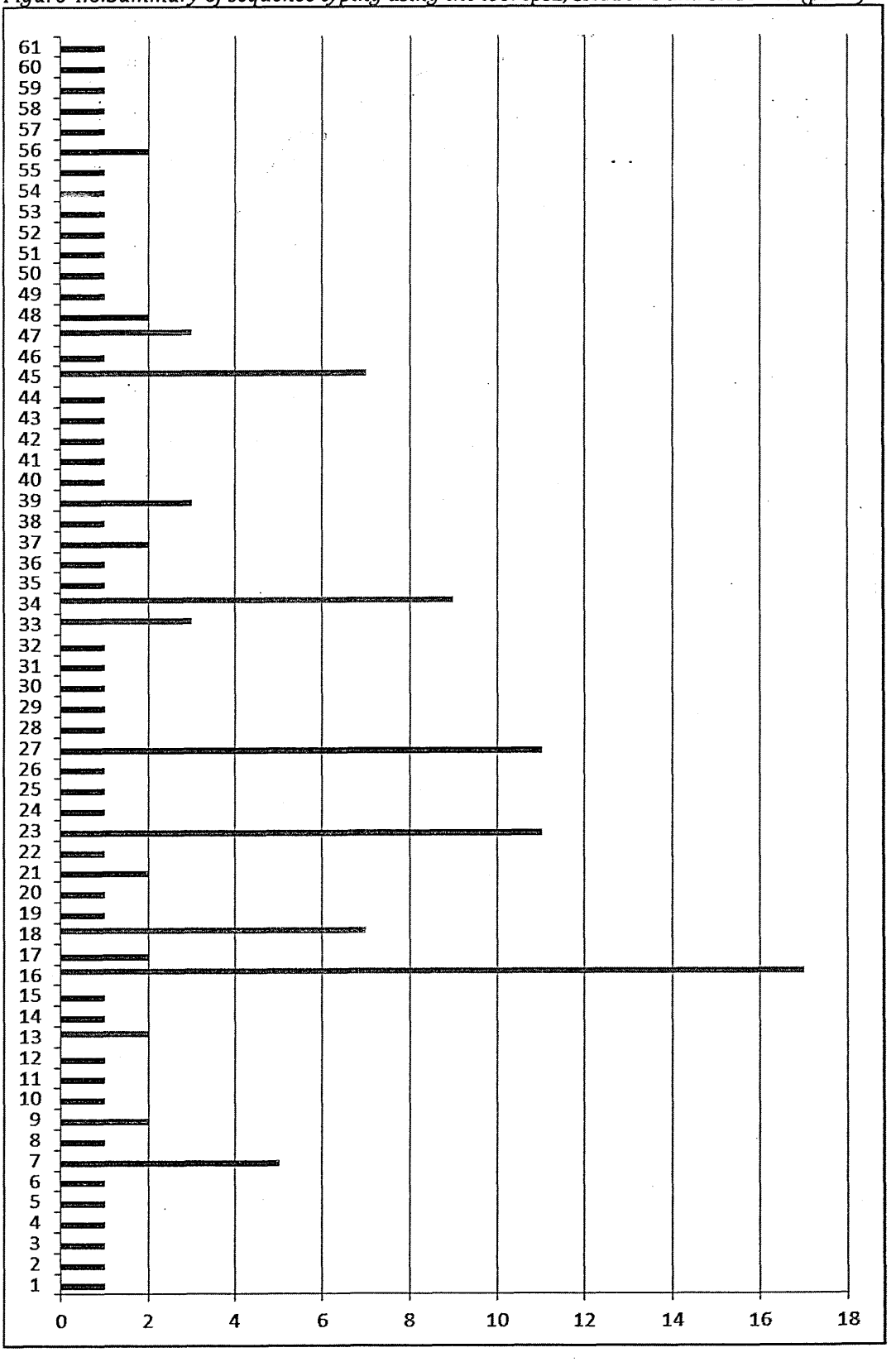
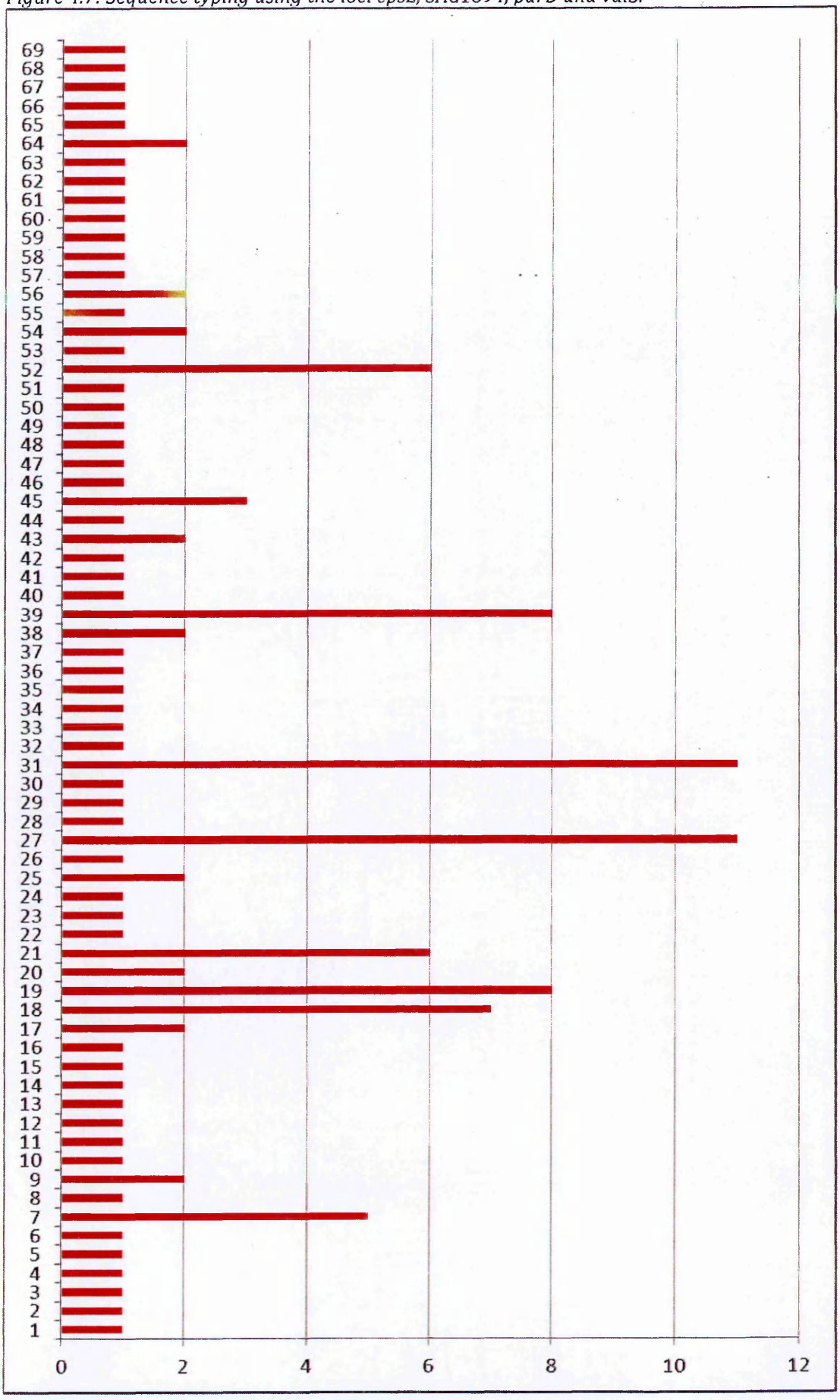


Figure 4.7: Sequence typing using the loci cpsL, SAG1894, purD and valS.



These results show that the addition of the *valS* loci generated an extra eight sequence types compared to the three gene profiling system. Of these eight sequence types, six sequence types of the three gene profiling system (ST-13, ST-18, ST-33, ST-34, ST-45 and ST-47) were split into one additional sequence type when *valS* was added. All of these new sequence types contained only one isolate. The largest sequence type, ST-16 (n=17) was split into three new sequence types (the 4 gene profiling sequencing sequence types are ST-17, ST-18 and ST-19) which may be useful. Since in total only 8 new sequence types were created the additional discriminatory power of adding *valS* is not sufficient to justify using a four gene profiling system.

4.2.3 Comparison of MLST to Novel Profiling Markers

The sequences from each loci of the MLST set and the 3 gene profiling method were concatenated separately and aligned. They were then analysed for the nucleotide identity, the number of unique allele types produced and the number of informative sites (Table 4.16). The alignments were also used to generate a maximum likelihood phylogenetic tree demonstrating the relationships of MLST (Figure 4.8) and 3 gene (Figure 4.9) sequence types. Also the isolates contained within each sequence type in each clade of the MLST and 3 gene profiling phylogenetic trees were compared to identify clades which were homologous between sequence typing methods.

Table 4.16: The discriminatory power, nucleotide identity and number of informative sites of each profiling scheme.

Scheme	Number of Loci	Total BP	Nucleotide Identity	Unique Allele Types	Number of Informative Sites*
MLST	7	3456	97.97	43	66
3 gene	3	1472	88.25	61	142

* The Number of informative sites corresponds to the number of phylogenetically informative columns in an alignment

Figure 4.8: A maximum likelihood tree indicating the relationships between MLST sequence types using the GTR model, optimised proportion of invariable sites and optimised gamma shape parameter. Highlighted is a potentially virulent clade.

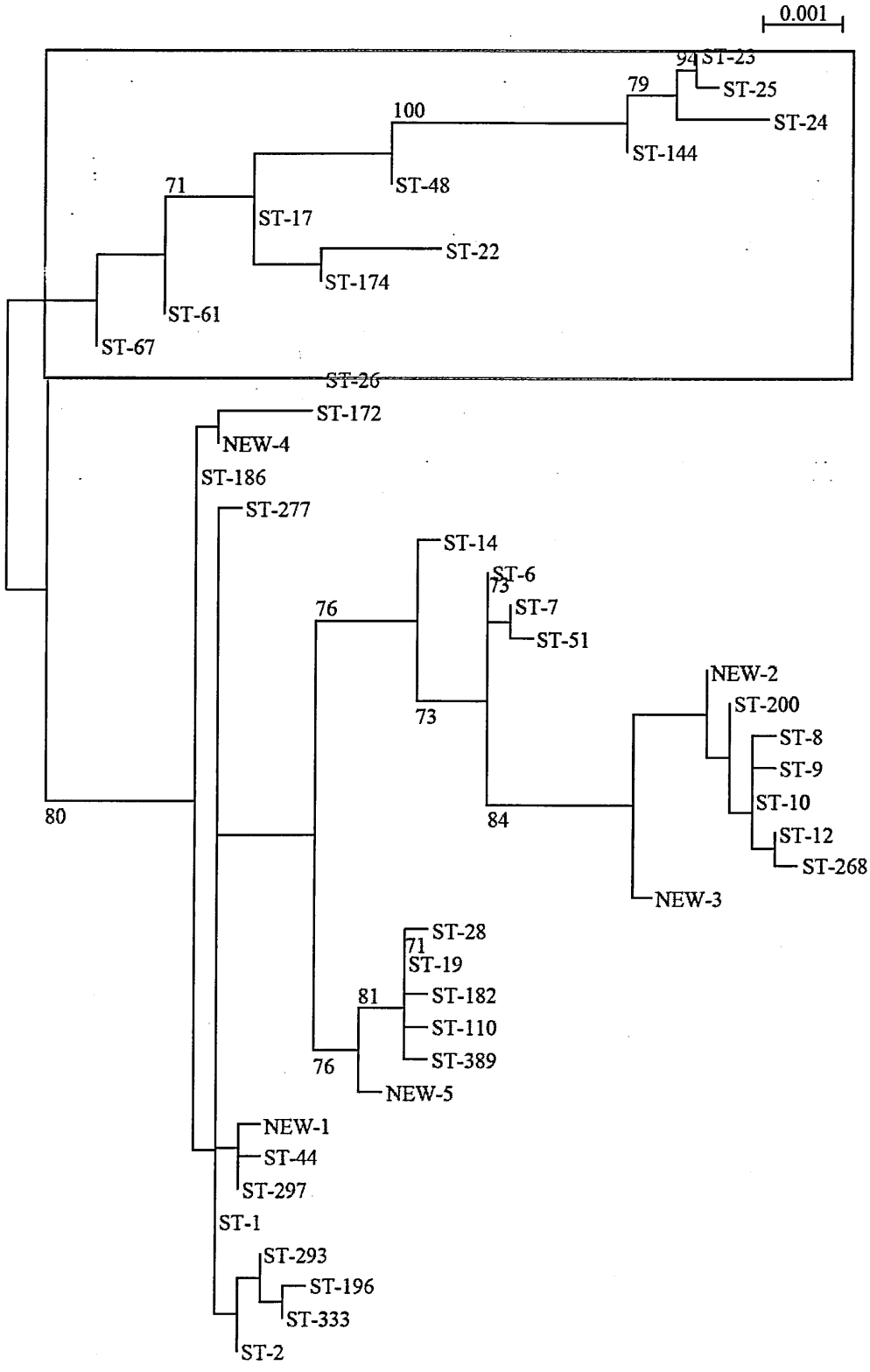
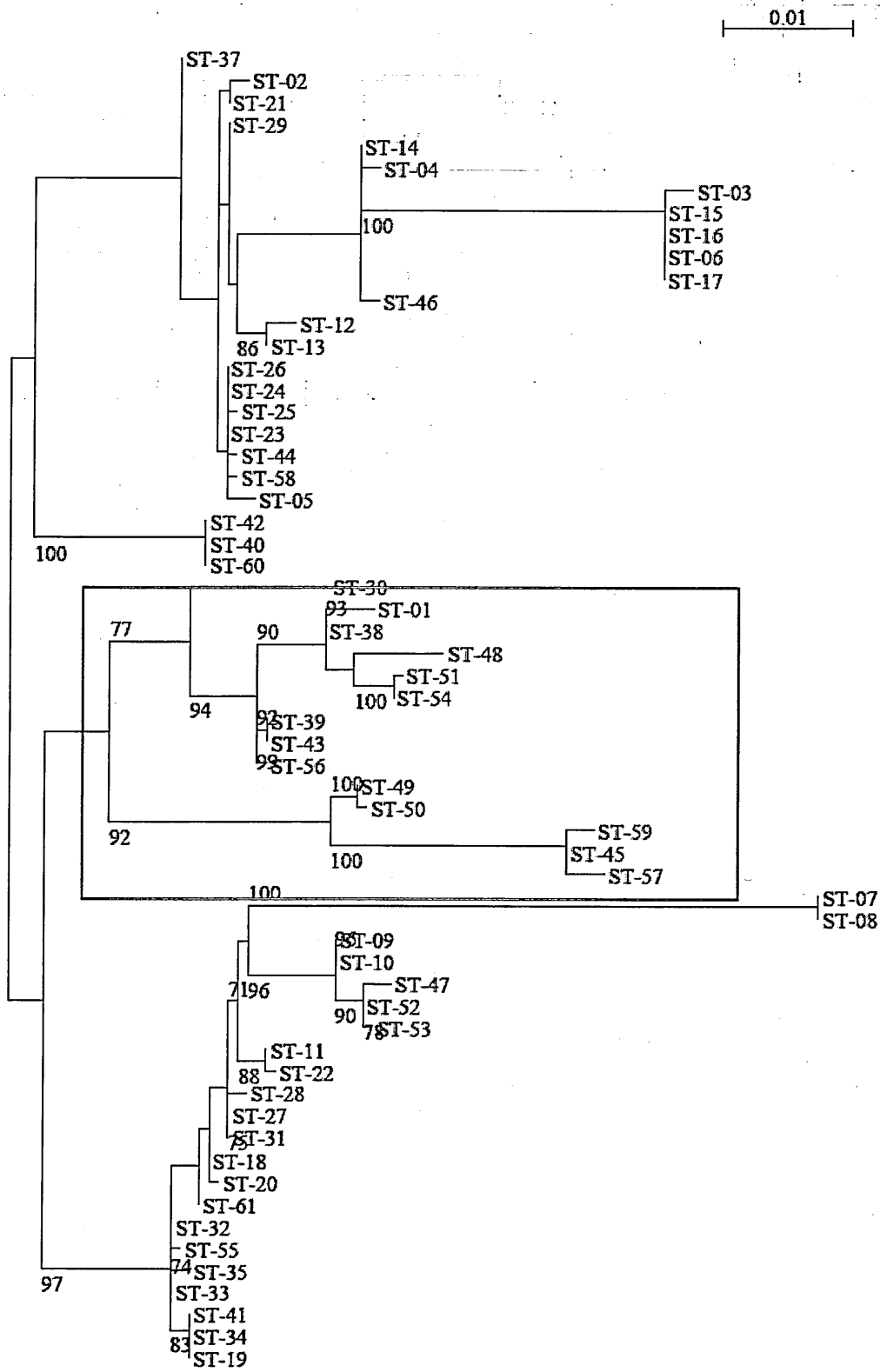


Figure 4.9: A maximum likelihood tree indicating the relationships between 3 gene profiling method sequence types using the GTR model, optimised proportion of invariable sites and optimised gamma shape parameter. Highlighted is a potentially virulent clade.



These results showed that the relationship between sequence types of the three gene profiling method (Figure 4.9) has stronger bootstrap support than the corresponding tree of the MLST set (Figure 4.8) when using the same maximum likelihood settings.

Also, comparing isolates found within clades showed a strong correlation between a clade in the 3 gene profiling tree to a clade in the MLST tree that contains sequence types previously identified as virulent (highlighted clades in figures 4.8 and 4.9) as these clades share 83% of isolates. Therefore, the three gene profiling system was similar to the MLST scheme in the ability to identify virulent strains.

The three gene sequencing method represents the phylogeny much better with the two main clades in the tree mirroring the two halves of the ANI tree. More specifically, the relationships between pairs of most sequenced genomes seems correct, A909 and H36B are indicated as being closely related according to the ANI and MLST trees, as are the strains 515 and NEM316. Using this method the strains 2603V/R, 18RS21 and CJB111 are all closely related and found in the same sub-clade of clade 2 and although the 3 gene sequencing tree showed that 2603V/R and CJB111 are more closely related than 2603V/R and 18RS21, which is shown in the ANI tree, the relationship is still closer than that indicated by the MLST tree. Interestingly, although COH1 is indicated as being related to 2603V/R, 18RS21 and CJB111 by one common ancestor, in the 3 gene tree it is found in a separate sub clade of clade 2. Despite this, the 3 gene tree more closely models the relationships between the sequenced genomes according to a tree based on the core genome. It is also worth considering that the core genome identified for the 8 genomes may not be accurate due to missing coding regions from the partially sequenced genomes. However, in the absence of other data this method has to be considered as a reference for the accuracy of other data.

The MLST scheme and the 3 gene profiling method were both analysed using the START2 (Sequence Type Analysis and Recombination Tests version 2) to determine if there is linkage disequilibrium between the loci. This was performed because proving linkage disequilibrium

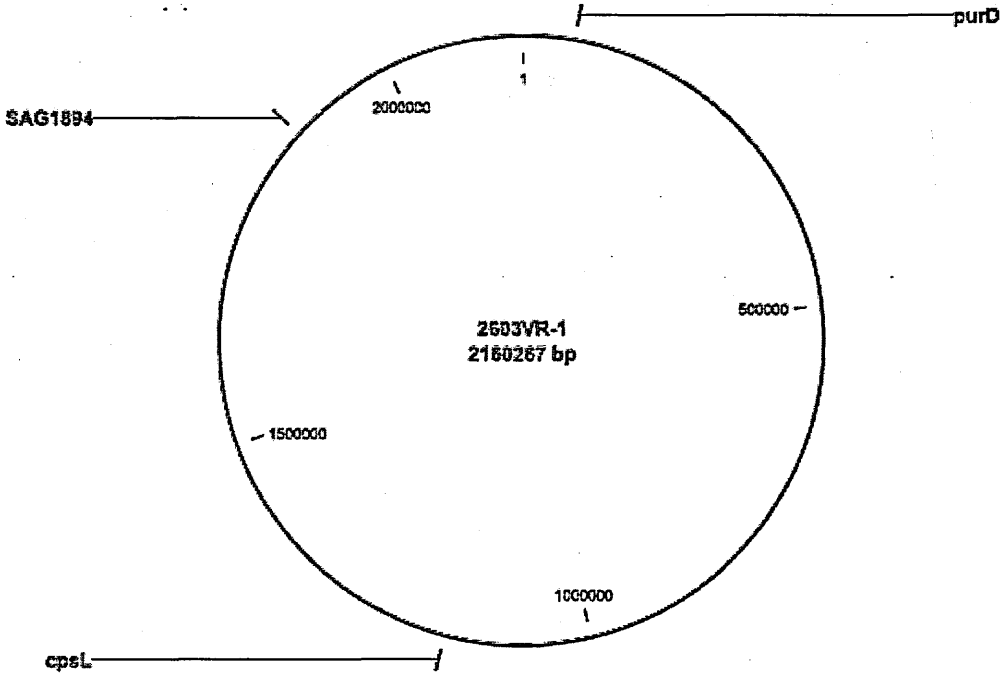
would prove that the loci of each typing method are linked and therefore evolving at the same rate, i.e. if an allele type from one loci is more likely to be associated with a specific allele type from another loci, and these loci are sufficiently far apart on the chromosome to not be affected by chromosomal rearrangement then it is likely that that hypothetical pair of loci are evolving at the same rate relative to other lineages. To test for linkage disequilibrium a classical (Maynard-Smith) and Standardised (Haubold) index of association test was performed on the allelic profiles of all available sequence types and for both profiling methods. Both index of association tests demonstrated a significant probability of linkage disequilibrium. This suggested that both profiling schemes were modelling evolution and not random change. This was further supported by the ratio of non-synomous to synomous mutation rates (dN/dS) also calculated using START2 in the selected loci. All dN/dS ratios were all <1 which suggested purifying selection.

The combination of the bioinformatic method used to select the targets of the three gene system has created a sequence typing system that improves on the discriminatory power of the MLST scheme using less targets. The proven linkage disequilibrium and the fact that all loci are under purifying evolutionary pressure suggests that the three genome method is modelling whole genome evolution. The MLST scheme however, is already acknowledged as modelling evolution over longer terms, as shown in this study by lower rates of variation in the MLST loci. Therefore, whilst both methods model evolution, and will usually agree on which isolates are related to each other, the new three gene profiling method developed here is more likely to be modelling recent evolution making it more useful as an epidemiological method for either surveillance or outbreak studies whereas the MLST scheme maybe better for identifying specific lineages (for example differentiating between bovine and human isolates (16)).

To summarise, the three gene profiling set provides increased discriminatory power. The level of nucleotide variation and the number of informative sites between orthologs of the loci of the three gene profiling method were significantly higher when compared to the MLST scheme and the number of unique allele types was also significantly higher (Figure 4.6). This method of target selection should identify the targets more representative of whole genome evolution and the

three targets selected were spread throughout the genome of 2603VR and therefore sequence types should not be biased by recombination events (Figure 4.10).

Figure 4.10: The position of each loci of the three genome profiling method on the 2603V/R chromosome



4.3 Discussion of the Analysis of the Core Genome

There have already been a substantial number of studies to learn more about the population structure of GBS using MLST in different countries. These include the United Kingdom (103,104), the United States (17), Sweden (141), Portugal (156), France (125), Israel (152), the Central African Republic and Senegal (23). In all of these studies the sequence types 1, 17, 19 and 23 appear prominently, although with some variation. How likely would it be that GBS isolates from four out of seven continents, or even from within one country, were genetically identical as indicated by the MLST results produced? This would imply that either GBS is a very old and very slowly evolving pathogen that grew alongside human populations or that being primarily a gastrointestinal bacterium that most of the world is consuming a food source infected with bacteria from the same source. The latter is not very likely. The more probable explanation is of course that the MLST scheme lacks resolution and is not differentiating between genomically different GBS isolates and suggests that the housekeeping loci of the scheme are very well adapted to their function and are under strong positive selection. These issues have been previously observed in other organisms (83) and are partly why whole genome sequencing, which gives higher resolution, is becoming so popular.

In defence of sequence typing using a limited number of core genes there are existing schemes that produce better discrimination and can be used to answer epidemiological questions, if you compare the GBS MLST scheme with the *Burkholderia pseudomallei* MLST scheme you see a more realistic distribution of sequence types worldwide. The sequence types generated from the *B. pseudomallei* MLST scheme have been plotted onto a world map using MLST maps (http://bpseudomallei.mlst.net/earth/maps/). Each sequence type is confined to one country with the exception of ST-40 which is distributed across a number of countries. However even with this sequence type it appears there is a pattern with the sequence type originating in China, spreading outwards and into the Middle East and into Europe. Obviously this is not a detailed look into *B. pseudomallei* epidemiology but the point here is that for the vast majority of *B. pseudomallei*

isolates a sequence type generated using MLST could determine the country of origin of the isolate. The data generated from the GBS MLST scheme does not allow this analysis.

The first aim of this project was to analyse the core genome to identify genes that are 1)more discriminatory than that of the GBS MLST scheme. Since a more discriminatory scheme alone would increase the probability of identifying different groups of GBS, 2) more phylogenetically representative of the whole genome. Thus creating a sequence typing method that more accurately reflects the relationships between GBS clinical isolates. This was achieved by adapting a method used by Konstantinids et al. (119) which has previously been used to select genomic markers for *E. coli* and the *Salmonella*, *Burkholderia*, and *Shewanella* groups.

Calculating the core genome of two separate datasets of three and eight genomes showed when each new genome is added the number of core genomes falls, which is confirmed by the findings of Lefébure et al (132). The proportion of core genes across the genome was also confirmed using, the Panseq tool (124) and the Multi-Genome Homology Comparison tool (JCVI), which showed that the level of core genes produced was similar but not identical. This is unsurprising considering that the JCVI method uses the criteria that 50% of the columns in the alignment must be shared. This is less stringent than the settings used here. The Panseq tool predicts regions of the genome that are core and may or may not be coding regions. Coding regions are then identified using GLIMMER which may pick out coding regions incorrectly. For example the 2603V/R genome has 2126 protein coding genes according to the NCBI database and 2145 according to GLIMMER. Additionally, the core genome generated by the Reciprocal BLAST method was confirmed by other studies that have identified the core genome of GBS. Lefébure et al (132). used a similar method except that instead of performing BLAST using length and homology criteria, they used an e value of 1e-5 and an inflation parameter of 1.5. Then when the alignments were created using clustalW to impose a cut off that meant any alignments that share less than 50% of conserved sites were rejected from the core genome. In other words they calculated a broad core genome and then removed elements rather than creating a stringent core to begin with and analysing it. Tettelin et al (231) also calculated the core genome of GBS using a radically

different method than the one used here. Their approach used three separate analyses, a Smith and Waterman protein search on all of the predicted proteins by using the SSEARCH program, a DNA search of all predicted ORFs of a strain against the complete DNA sequence of the other strain using the FASTA program and a translated protein search of all of the predicted proteins of a strain against the complete DNA sequence of the other strain using the TFASTY program. This resulted in a larger core genome than identified in this study. This is not surprising since they used three separate methods and if any one of these methods identified a gene as core it was added to the list which may have overestimated the core genome. Particularly as using translated protein searches as if the majority of nucleotide differences do not affect the amino acid sequence then this may severely underestimate diversity. In conclusion, the core genome extracted here using the Reciprocal BLAST method is supported by multiple methods and published literature.

Analysis of what is contained within the core genome using the COG category system revealed some interesting results which may have affected previous analysis of the core genome. The data showed that the core genome of the 3 genome dataset had 2 categories (nucleotide transport and metabolism and the coenzyme transport and metabolism) that had 100% of its genes in the 3 genome core, however a sharp reduction in the number of these genes was observed in these 2 categories in the 8 genome dataset. This may suggest that the core genome is being underestimated due to core genes being missing from the whole genome shotgun sequences, since both of these categories are most likely essential for the survival of the organism. For example the *atr* gene which encodes an amino acid transporter and is used in the MLST scheme is not completely present in the 18RS21 genome but there are no examples of this gene being missing in any isolates profiled by MLST. Because of this and because it is unlikely that amino acid transport is a non-essential function, it is therefore unlikely that this gene, and others like it, are not present in all genomes. It is possible that some genomes have separate genetic mechanisms for performing the same function. However, such a large number of genes believed to be essential to the survival of the organism not being present suggests that the likely explanation is that these genes are missing from the partially sequenced genomes. This may be possible because

both of these categories contain genes encoding transport proteins which are known to contain repeat regions which can be difficult to assemble into contigs depending on the genome sequencing method employed. If the level of decrease in the size of the core genome as new sequenced genomes are added is overestimated it would have implications for previous work demonstrating that GBS has an open pan-genome (132,231) as the calculation of the rate of decrease in the size of the core genome would be incorrect. It would also suggest that the estimation of the pan-genome may be overestimated as any genes added to the pan-genome based on these analyses could actually be part of the core genome. In fact, Tettelin et al. (231) suggest that the findings from their work may challenge the findings of Konstantinidis and Tiedje (120) by suggesting that rather than limited variation, bacterial species may be much more variable than suggested because of increased variation in the pan-genome. However, if the analysis of the core genome was unreliable this would bring that hypothesis into doubt. Yet, making this conclusion may be somewhat premature considering this was based on looking at different species of bacteria and it is well known that each species may evolve differently and be more or less prone to gene acquisition and loss. Secondly, the analysis by Tettelin et al. used partially sequenced whole genome shotgun sequences which based on this analysis would underestimate the core genome. To confirm this finding would require either completion of the 5 wgs genomes or analysis of additional genomes which was not possible in this project but it is an interesting finding which should be considered in analysis of these core genomes and potentially in further work. Fortunately, since this project was completed a large number of new GBS genomes have been completed, for example NCBI now shows 143 genomes with data published in the wgs database and an additional 105 with data in the Short Read Archive (SRA) The majority of these coming from the JCVI which is currently running a project on evolutionary genomics and the population structure of pathogenic *streptococci* and in addition to these genomes is also due to publish data on a further 67 genomes.

The ANI calculated using the 3 genome and 8 genome datasets showed that, on average, the core genomes are remarkably similar and are well above the 94% ANI value that is viewed as

equivalent to the traditional 70% DNA-DNA reassociation standard of the current species definition (120). It is also significantly higher than ANI values calculated for the four bacterial groups studied previously, which is understandable as a group would be expected to be more variable than a species. The ANI value is also higher than that of the *E. coli* genomes. It has been suggested that they have low ANI values and lower percentages of core genes because they cover a larger number of environmental niches and therefore require more variation than a pathogen that is restricted to limited hosts. Previous work also showed that the larger the genome of an organism the lower the ANI which means GBS with a relatively small genome would be expected to have a higher ANI as was shown here. This suggests that the majority of variation in GBS comes from recombination and that GBS is a pathogen more highly adapted to its surroundings and therefore evolving at a slower rate than organisms that are more promiscuous in their environmental niches. It also suggests that selecting housekeeping genes from the core genome is likely to give a less discriminatory typing scheme since these genes are likely to be less variable than this already highly similar average, as shown with the MLST genes, and traditionally are selected for molecular typing schemes for this reason (103,146).

The ANI was also used to generate a tree indicating the relationships of the genome sequences using the ANI meaning the tree shows how the core genome is evolving. This showed that genetic relationships are not related to serotype as has previously been observed (150,219). A similar method has since been used to perform phylogenetic analysis on MRSA isolates from a Thai hospital sequenced using the Illumina genome analyser (83) which showed that using core genome SNPs differentiated a collection of ST-239 MRSA isolates on the basis of their country of origin and even was able to identify differences between isolates from patients from different wards of the same hospital suggesting that using core genome SNPs provides a very accurate method for phylogenetics and typing. An alternate method for typing using the whole genome would be pan-genome typing (79) which has been shown to differentiate between the sequenced genomes of six bacterial species. However, using this method to analyse the core genome would give no phylogenetic information to allow target selection.

Since the target selection process was performed on two datasets, one of which contained three genomes, an addition to the Konstantinidis method (119) was required to further discriminate them since mathematically every possible combination of ranks can only give seven categories which needed to be separated out for target selection. This was done by measuring the distance between the ANI and the distance value which allowed target selection 1) based on the distance value of the target relative to the ANI or 2) to select highly variable targets that are still evolving in the same way as the whole genome. This could be useful for design of profiling schemes with only a limited number of published sequences although technological advances have made this less of an issue.

Targets were not selected solely on the basis of the computational analysis as in Konstantinidis's work, rather a selection of targets were chosen and sequenced to determine the variation at the locus level and the number of unique allele types produced and then combined to determine how discriminatory the targets would be when used together. This showed that the most variable targets do not make the most discriminatory profiling scheme. For example, the four most variable targets with a Kendal's Tau score of 1 selected from the three genome dataset produced even less unique sequence types than the MLST scheme. This is probably because these genes are all part of the same biosynthesis pathway (188) and are therefore under the same selective pressures and are evolving at the same rate, although quicker than the MLST housekeeping genes. Therefore, another variation to the Konstantinidis method (119) was developed to determine the most variable combination of sequences, if the targets were selected solely on the basis of relationship to the ANI as in the Konstantinidis then the sequence typing system produced would have been less discriminatory than the MLST scheme.

Interestingly, after three targets the number of unique allele types generated increased at a slower rate until adding an extra target did not increase the number of unique sequence types at all after 8 genes were added to the sequence typing method. This suggests adding further targets would not significantly improve discrimination. This is in contrast to current methods which involve expanding the number of housekeeping genes in the MLST scheme as used by Sorrensen

et al. (219), which would suggest that the level of variation in these genes in insufficient for molecular typing.

Reliable interpretations of phylogeny in this strain collection are difficult due to a lack of strain information. It is possible however to show how each scheme differentiates between and indicates the relationships between the genome sequenced isolates by comparing the trees of each method to the tree of the whole genomes as determined by the ANI which indicates the relationships between the core genomes.

The tree based on the MLST data showed that the strains COH1, NEM316 and 515 are located in the same clade, the whole genome tree indicates that 515 and NEM316 are located together but COH1 is more distantly related, meaning MLST is assigning the relationship between COH1 and NEM316 and 515 incorrectly. The isolates A909, H36B and 2603V/R are also clustered together according to the MLST tree, in the whole genome tree A909 and H36B are closely related but 2603V/R is much more distantly related. Additionally, the isolate 18RS21 is not particularly closely related to any of these groups but is most closely related to the isolates A909 and H36B according to the MLST tree but are distantly related according to the whole genome tree. These data suggests that the MLST scheme may not be correctly measuring the relationships between isolates and is therefore misrepresenting the evolutionary relationship between GBS isolates.

Phylogenetic analysis using these three targets showed that bootstrap support was stronger for the three gene tree than the MLST tree. This is most likely because there are more variable sites in the three genes than in the seven MLST loci, resulting in more phylogenetic information from which to draw conclusions about relationships between isolates. These data shows that the three gene profiling method is more discriminatory and generates trees that are more highly supported than those generated from the MLST scheme. However, to prove conclusively that this scheme is performing better than the MLST scheme it would need to be tested on a larger and more varied strain set, ideally with clinical isolates from locations outside the UK to test the theory that better targets would discriminate between an isolates country of origin.

This work has shown that three genes selected using a bioinformatic approach results in sequence typing with superior resolution and accuracy compared to the seven genes of the MLST scheme. This implies that the method used for selecting targets for sequence typing is more important that the number of targets sequenced. That is, as sequencing technology continues to improve and the cost of sequencing decreases while the throughput increases, some thought about the markers that are being used in sequence typing methods provides a greater improvement in accuracy and resolution than a brute force method of sequencing more targets. It has also shown that using a bioinformatic method alone would not produce the most efficient typing system and a bioinformatic approach must be combined with laboratory testing using clinical isolates, at least until there is a higher number of sequenced genomes. Using fewer targets also has implications for epidemiological studies and research projects as a higher number of isolates can be studied in a shorter time frame with much reduced costs and less analysis required.

Chapter 5

Analysis of the MNR repeats for Profiling

5.0 Bioinformatic Analysis of MNRs for Profiling

5.1 Aims and Objectives

In addition to developing an accurate method to assess the phylogeny of GBS clinical isolates this section looked at methods to determine potential differences in virulence between different GBS clinical isolates. To do this the content of GBS genomes was assessed for virulence factors, since these were shown to be homologous between the sequenced genomes the level of Mononucleotide Repeats (MNRs) within core genes was assessed with a focus on repeats found in known virulence genes as these were shown to be involved in genomic regulation through slipped strand mispairing creating truncated gene products (133). Additionally, the presence and abundance of MNRs were assessed in non-coding areas of DNA around known virulence factors as these have been shown to have an effect on genomic regulation (165,242) and have also been shown to be good targets for strain profiling (42). The targets generated were then assessed as profiling markers and compared to both the existing MLST scheme and the previously discussed three gene profiling method. Final analysis of all profiling markers was also performed.

5.1.1 MNRs Within Coding DNA of Core Genes

It has previously been shown that MNR repeats are overrepresented in the first 10% of coding DNA sequences (171) and it is believed that repeats act as a regulator of transcription through slipped strand mispairing causing a frame shift that brings an out of frame stop codon into frame. When this occurs earlier in transcription it is better for the organism since it uses less energy in creating a non-functioning truncated RNA.

Using a custom perl script (Appendix 9.9.4) each core gene from the 8 genome dataset was scanned to identify all MNR repeat regions greater than 6bp in length and used to indicate the level of homology of each tract in other words, each MNR tract was scored on how many core gene orthologs it was found in. This data was used to identify genes that had MNR repeats in the first 10% of the gene and to assess any difference in length and/or presence of repeats in core

genes. Genes that have previously been identified as virulence factors were considered the most important because the majority of GBS virulence factors are present in all strains of the organism therefore differences in virulence between strains is most likely due to regulation of transcription.

In total 927/1179 genes contained at least one MNR repeat with the majority (83.3% of MNR repeats) being homologous across all genomes, i.e. the MNR repeat region was present at the same location and was the same length and composition in all 8 core gene orthologs. The position by percentage of gene length of each core gene was then calculated and only genes with MNR repeat regions in the first 10% of the gene were analysed. Of the 927 genes containing MNR repeats >6bp in length 325 genes were found to contain at least one MNR in the first 10% of the coding sequence, meaning a higher than average level of MNR repeats are found in the first 10% of genes.

Genes with MNR repeat regions in the first 10% were correlated to known virulence factors since regulation of these genes would make the biggest impact in regulation of virulence in the organism. Surprisingly only 6 virulence genes were identified with MNR repeat regions in the first 10% of the gene and all of these repeats were homologous across all sequenced genomes (Table 5.1).

Table 5.1: Virulence factors with repeats in the first 10% of the gene

Loci	Product	Repeat	Position	Length of Repeat	Number of genomes with homologous repeat	Length of Gene
SAG0667	cylA protein	AAAAAAA	26	7	8	929
SAG0668	cylB protein	TTTTTT	65	6	8	878
SAG1159	Capsule	AAAAAAAA	3	8	8	629
SAG1161	Capsule	AAAAAAA	57	7	8	1025
SAG1172	Capsule	AAAAAAAA	41	8	8	689
SAG1173	Capsule	AAAAAAA	50	7	8	692

To test the hypothesis that genes selected containing MNR regions may be involved in slipped strand mispairing the number of NMR containing genes found in each COG category was compared. If MNRs were repeatedly found in genes with essential functions, for example categories J and K (corresponding to genes involved in translation and transcription respectively) compared to non-essential genes such as category V (defense mechanisms) or genes with no known COG category then this method may not be predicting repeat regions that are involved in genomic regulation, suggesting the repeats were present for structural reasons (Table 5.2).

Table 5.2: COG categories of all genes, all MNR containing genes and genes containing MNRs in the first 10% of the gene.

COG	Genes in COG (2603V/R)	Genes in COG (2603V/R, %)	Genes in COG (all MNR)	Genes in COG (all MNR, %)	Genes In COG (MNR in first 10%)	Genes In COG (MNR in first 10%, %)
J	152	6.47	87	8.64	30	8.45
K	160	6.81	76	7.55	26	7.32
L	144	6.13	48	4.77	21	5.92
D	24	1.02	13	1.29	2	0.56
V	45	1.92	23	2.28	11	3.10
T	74	3.15	41	4.07	14	3.94
M	114	4.86	51	5.06	19	5.35
N	8	0.34	4	0.40	1	0.28
U	27	1.15	12	1.19	5	1.41
O	57	2.43	31	3.08	11	3.10
C	61	2.60	30	2.98	12	3.38
G	153	6.52	77	7.65	24	6.76
E	160	6.81	90	8.94	32	9.01
F	74	3.15	38	3.77	13	3.66
H	51	2.17	37	3.67	9	2.54
I	51	2.17	28	2.78	8	2.25
P	109	4.64	49	4.87	19	5.35
Q	27	1.15	9	0.89	2	0.56
R	247	10.52	100	9.93	37	10.42
S	166	7.07	65	6.45	21	5.92
-	444	18.91	97	9.63	37	10.42

Table 5.2 shows that the majority of COG categories were present at approximately the same proportion in both all MNR repeat containing genes and genes containing MNR repeats in the first 10% with the only exception being genes not found in any COG category. To show how similar the proportion of genes predicted in both MNR datasets were the Kendal Rank correlation coefficient was calculated for the total genetic content of 2603V/R and all MNR containing genes and between the total genetic content of 2603V/R and genes containing MNRs in the first 10% of the gene. This showed a correlation of 0.85 and 0.90 respectively suggesting a strong correlation and that MNR repeat regions were found across a wide variety of genes with no particular preference for genes likely to undergo regulation by slipped strand mispairing.

Analysis of COG categories of genes and the lack of homology found in MNR repeat regions within genes suggests that either this method was not reliable for identifying genes regulated via slipped strand mispairing. It is also possible that the collection of sequenced isolates is biased in favour of virulent isolates. For this method to be useful genome sequences of non-virulent GBS isolates would be required. Therefore, these data were not investigated further.

5.1.2 MNRs Within Non-Coding DNA

Mono-Nucleotide Repeats (MNRs) within non-coding DNA have previously been identified as a source of hyper-variability with higher rates of variability observed within MNR tracts and in the flanking sequences in *Escherichia coli* when compared to the housekeeping genes of the *E. coli* MLST scheme (42). MNRs in non-coding DNA have also been identified as potential regulatory site for virulence genes (184). Therefore, non-coding DNA sequences containing MNRs were identified as potential profiling markers.

To identify MNR tracts within the non-coding genome two versions of the genome, one positive strand and one negative strand were created that do not contain DNA coding regions for each of the three fully sequenced genomes. The strand each coding loci was on was taken from the NCBI protein tables. The positive and negative strand coding loci were then separated and this analysis showed that the genomes 2603V/R, A909 and NEM316 had 47.8%, 50.8% and 48.3% of their coding loci on the positive strand respectively.

Using a custom Perl script (appendix 9.9.1) the coding loci were trimmed by 20bp at the 5' and 3' ends to allow replacement of coding sequences that overlapped. This gave an area to design primers. These trimmed coding loci were then removed from the chromosomal DNA and replaced with the locus tag of the coding DNA. The result was two files containing non-coding regions on the positive and negative strands. This resulted in some of the non-coding regions containing coding DNA from the opposite strand. These were easily identified by size and by BLASTing selected targets against the genome to confirm true non-coding regions and these were removed from the analysis.

Another custom Perl script (appendix 9.9.3) was then used to identify any NMR tracts >6bp in each non-coding region and to return the 5' and 3' locus tags, the length of the repeat and the composition and the length of the non-coding region. Across all genomes this showed a total of 539, 511 and 534 non-coding regions that contained MNRs for the strains 2603V/R, A909 NEM316 respectively. However, since these lists include "pseudo non-coding" regions and the targets that

were chosen for this method were designed for short sequencing, only non-coding regions between 200-300bp were considered in any further analysis. A summary of the MNR repeat containing regions is shown in Table 5.3, and a summary of all MNRs within non coding DNA is shown in Table 5.4.

Table 5.3: The number of MNR repeats in non-coding DNA between 200-300bp.

Nucleotide	Repeat Length	Frequencey (2603V/R)	Frequencey (A909)	Frequencey (NEM316)
A	6	25	29	27
	7	9	9	13
	8	1	0	3
	9	0	0	0
G	6	1	1	0
	7	0	0	0
	8	0	0	0
	9	0	0	0
C	6	0	0	0
	7	0	0	0
	8	0	0	0
	9	0	0	0
T	6	21	21	20
	7	11	7	11
	8	1	1	2
	9	1	1	1
	Total	70	69	77

These potential profiling markers were then correlated to known virulence factors extracted from the Virulence Factor Database (31). It is probable that regions of the genome containing virulence factors were more variable due to immune evasion factors and also that MNR regions in non-coding DNA have been linked to regulation of virulence factors (184). This analysis showed that in the genomes 2603V/R, A909 and NEM316 there were five, two and three 200-300bp non-coding loci which had a virulence factor (VF) at either the 5', 3' or both ends respectively (Table 5.5).

Table 5.4: A summary of all MNRs found in non-coding DNA

	2603VR	A909	NEM316
Number of Coding Regions	2121	1996	2094
Total Non-Coding Regions	1057	995	1039
Non-Coding with MNRs	539	511	534
Non-Coding with MNRs and VF at 5' and/or 3'	17	22	15
Non-Coding with MNRs and VF at 5' and/or 3' 200 ≤ 300bp	5	2	3

Each of the ten potential profiling markers were searched using BLAST against each other to remove homologs and to identify loci that were present in all 3 sequenced genomes. Four loci were identified that met the above criteria (SAG0032-SAG0033, SAG0649-SAG650, SAG1768-SAG1767 and SAG2044-SAG2043) and one locus that showed homology in all genomes but with an insertion sequence separating the region found in A909 in the 2603V/R genome (SAK_1320-SAK_1319) as shown in Table 5.5.

Table 5.5: Profiling targets selected from the non-coding regions of the three fully sequenced genomes.

5' Loci	Length	Strand	3' Loci	VF	Sequenced?	Contain MNR Repeats
SAG0032	246	+	SAG0033	5'	✓	✓
SAG0649	274	+	SAG0650	5'	✓	✓
SAG1768	208	-	SAG1767	5'	✓	✓
SAG2044	219	-	SAG2043	3'	×	✓
SAK_1320	235	-	SAK_1319	5'/3'	✓	✓
SAG0106	272	+	SAG0107	No	✓	×
SAG0043	280	+	SAG0044	No	✓	×
SAG2143	228	-	SAG2142	No	✓	×

From these loci, three were selected for sequencing and further analysis (Non coding regions between SAG0032-SAG0033, SAG0649-SAG0650, SAG1768-SAG1767). These were compared to three randomly selected non-coding sections of DNA that do not contain MNR repeats (Table 5.5). The intragenic region between SAK_1320 and SAK_1319 was also selected for sequencing because firstly, a MNR repeat region was found in two out of three sequenced genomes and secondly because of the presence of an insert in one genome which has previously been linked to increased

lmb binding which may play a role in enhancing virulence of GBS (2,220). Primers were designed ~100bp upstream of each of the selected targets to ensure the entire non-coding region was sequenced.

5.2 Sequence Analysis of MNRs for Profiling

5.2.1 Comparison of MNR and Non-MNR Non-Coding Loci

The selected non-coding loci containing MNR repeat regions and the regions without MNR repeats were sequenced for all 134 isolates in the strain collection. They were analysed by allele typing of each loci, by assessing the level of nucleotide variation between orthologs of each target and by sequence typing selected combinations of loci.

The number of unique allele types and the level of nucleotide variation are shown in Table 5.6 and it was found that the levels of nucleotide variation and the number of unique allele types were relatively consistent across targets containing MNRs and those without. The SAG1768-SAG1767 intragenic region splits the strains into the highest number of unique allele types (n = 17) with a variation rate of 1.02 SNPs/100bp compared to the SAG0043-SAG0044 intragenic region which splits the same strains into 12 unique sequence types but has a level nucleotide variation almost 4 times higher (3.95 SNPs/100bp). This suggests that MNR repeat containing regions are better at discriminating sequence types than similarly variable regions of non-coding DNA that do not contain MNR repeat regions. The allele types were used to place the strains into unique sequence types and the ability of MNR containing loci with non MNR containing loci to cluster the tested isolates was analysed. The sequence typing using both sets of targets is shown in Appendix 9.5 and 9.6 and the distribution of isolates across each sequence type is shown in figures 5.1 and 5.2 respectively.

Table 5.6: The number of unique allele types and nucleotide variation of the selected non-coding loci. In the unique allele types column numbers in brackets indicate sequence types that are genomic inserts.

Loci	Source	Length	Unique Allele Types	SNPs/loci	SNPs/100bp
SAG0032-SAG0033	Non-Coding + MNR	246	8	1.89	0.77
SAG0649-SAG0650	Non-Coding + MNR	274	3	0.38	0.14
SAG1768-SAG1767	Non-Coding + MNR	208	14 (3)	2.13	1.02
SAK_1320-SAK_1319	Non-Coding + MNR	235	7 (2)	1.44	0.61
SAG0043-SAG0044	Non-Coding	280	12	11.06	3.95
SAG0106-SAG0107	Non-Coding	272	6	0.64	0.24
SAG2143-SAG2142	Non-Coding	228	8	2.00	0.88

Figure 5.1: The distribution of sequence types for the non-coding loci containing MNRs

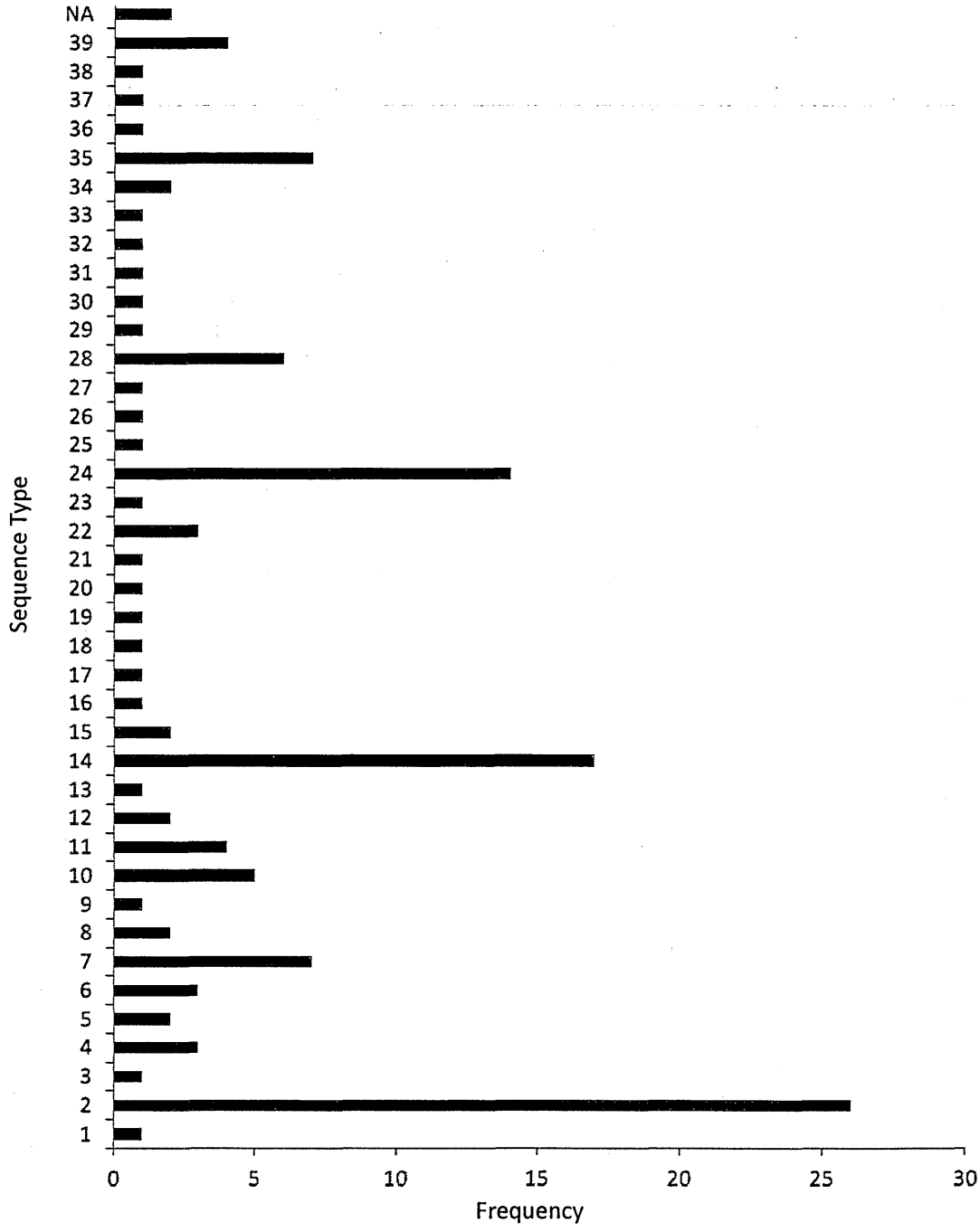
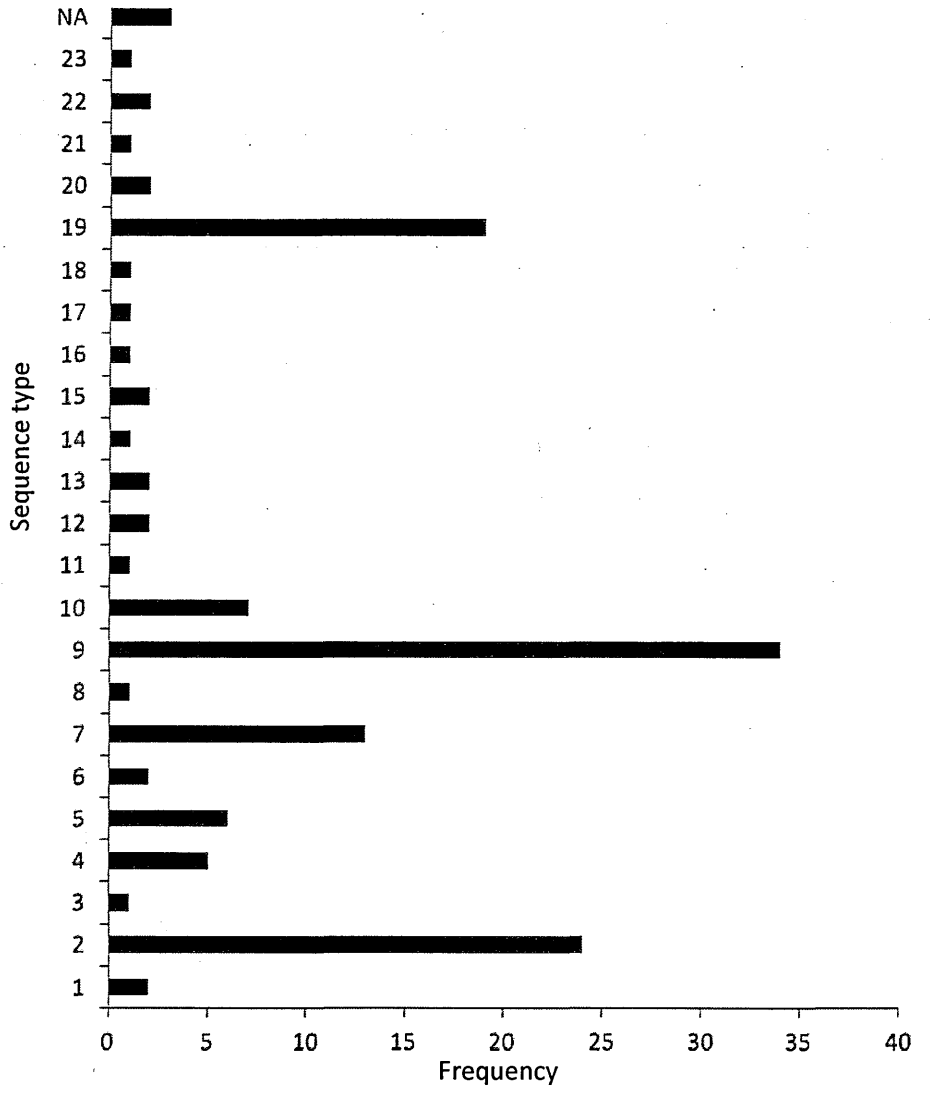


Figure 5.2: The distribution of sequence types for the non-coding loci not containing MNRs



Five loci from four strains failed to amplify. This is most likely because of a mutation in the primer binding region, because the genes are arranged differently in these strains or because the coding regions the primer sites are located in were not present.

In total, profiling using MNR repeat coding regions was able to sequence type 132/134 isolates and profiling using non-coding DNA that does not contain MNR repeats was able to sequence type 131/134 isolates. Using non-coding DNA that contains MNR repeat regions forms more unique sequence types (39) than using non-coding DNA that does not contain MNRs (23). The nucleotide identity of the MNR containing non-coding DNA targets was calculated and compared to the

nucleotide identity of the non-coding DNA targets not containing MNR regions. This analysis showed that the MNR containing DNA regions were on average more similar with 98.6% nucleotide identity compared to 97.7% nucleotide identity for non-coding regions not containing MNR repeat regions but was still able to split the isolates into more unique sequence types. Figures 5.1 and 5.2 show the distribution of sequence types and for both MNR containing non-coding DNA and non-coding DNA not containing MNRs the top three sequence types contain a large proportion of the isolates (42.5% and 57.5% respectively).

However, neither method using non-coding DNA created more unique sequence types than either the MLST scheme of the 3 gene profiling method which gave 43 and 61 unique sequence types respectively. Additionally, the top three sequence types of the MLST scheme and the 3 gene method contained a smaller percentage of the isolates (24.1% and 29.1% respectively).

5.3 Discussion of using MNRs for profiling

The second aim of this project was to identify regions of hyper-variability within the GBS genome to select variable markers to further discriminate GBS isolates. This was achieved by studying MNR tracts as they have previously been shown to be sources of variation due to replication errors by slipped strand mispairing, meaning these regions more likely to be prone to mutation than average (133). Slipped strand mispairing can occur due to MNRs in the genome forming transient mispaired regions during transcription leading to truncated RNA products and during replication DNA polymerase may undergo slippage at these locations resulting in either expansion or contraction of repeat units in the progeny generation (133). This has been shown that to increase the fitness of organisms in general (165,242) and specifically in pathogens including *Helicobacter pylori* (3,205), *Haemophilus influenza* (92) *Neisseria* spp (108,217) *N. meningitidis* (153) and *Campylobacter jejuni* (178,247).

MNR repeats were investigated by two approaches, firstly core genes were studied for the presence and homology of MNRs as slipped strand mispairing would cause that changes in the length of MNR tracts, causing a frame shift potentially leading to premature termination of the gene product by bringing an out of frame stop codon into frame. The position relative to the start of a gene was also considered because it has been shown that genes in the first 10% of a coding gene are more likely to be involved in slipped strand mispairing, as these repeat regions are likely to have evolved to limit the amount of energy the cell expends to create a non-functional transcript (171). Also because virulence factors are universal in all sequenced genomes there is likely to be regulatory mechanisms to prevent the exposure of surface proteins to the immune system. Due to the latter, this work focused specifically focused on virulence genes as any occurrences of slipped strand mispairing would have an effect on the virulence of the organism due to the regulation of expression of proteins that either come into contact with the immune system or affect the pathogenesis of the disease.

The second approach will look at MNR repeats in non-coding DNA as these have been shown to be involved in genomic regulation (140) and have also been shown to make good typing markers

due to the instability of the genome and therefore lead to increased variability of these regions (55) making them suited to profiling.

The bioinformatics analysis showed that the majority of core genes had an MNR repeats greater than 6bp in length but that only 27% of genes had MNR repeats in the first 10% of the gene sequence. Of these genes only eight encoded virulence factors, meaning the presence of MNR repeats in this organism are under-represented in these genes compared to the average.

In fact, when the classifications of genes containing MNR repeats were studied there was statistical evidence to suggest that there is no preference to genes that would be expected to be regulated by slipped strand mispairing. This was because there was no significant increase in the number of genes that have a MNR in the first 10% of the gene in genes expected to be regulated by slipped strand mispairing such as virulence genes and genes that code for proteins exposed to the host immune system and those not be expected to be regulated by slipped strand mispairing such as genes involved in transcription and translation.

A study by Janulczyk et al. in 2010 (97) examined the same issues except that they looked at all genes for evidence of slipped strand mispairing and not just on virulence genes and found only 39 genes that contained MNRs that were of a different length in at least one sequenced genome suggesting that these 39 genes are the only genes that are under regulation by slipped strand misspairing. However, just because there is homology between all other MNR regions in genes does not automatically mean that no other genes are under regulation by slipped strand misspairing. It is possible that they are not identified in the sequenced genomes since the sequencing used would only give the most prevalent length of MNR repeats when sequences are assembled into contigs. Considering this, it has been demonstrated that only 1-2% of GBS isolates (250) are non-hemolytic which means that if a culture was grown for sequencing a GBS genome, only 1-2% of the sequences generated for the MNR tract would have a frameshift mutation meaning when the sequences were assembled the MNR repeat region would be identified as having the non-frameshift length of nucleotides. The best way to establish if slipped strand

mispairing occurred would be to grow a culture and deep amplicon sequence the selected region to prove that a percentage of sequences have a different number of nucleotides at this repeat region. As the availability of wgs increases within the HPA, this may be an area for future work.

The second approach was to look at MNR containing regions of non-coding DNA for use as sequence typing markers. Non-coding DNA was selected for two reasons, firstly it may be involved in regulation of the upstream genes (184), which would allow profiling with a link to the phenotype of the organism. Secondly if these non-coding DNA are not involved in regulation it would imply that this DNA would be free of selective pressures and free to accumulate mutations making these regions a source of increased variability, particularly if these regions contain sequence repeats since these regions would still be prone to mutation due to replication errors (234). The majority of sequence typing methods that use targets in non-coding DNA use short tandem repeats and this method has been applied to *Mycoplasma genitalium* (145), *C. difficile* (259), *S. aureus* (206), *M. tuberculosis* and *M. bovis* (222). Using MNR containing non-coding region for selecting sequence typing targets has also been successfully applied to *E. coli* and this is the approach which was applied in this study (42).

Since these targets were designed for sequence typing, only non-coding regions between 200-300bp were considered. Within these limits, the three fully sequenced genomes on average had only 72 non-coding regions containing MNRs and only 3.3 on average between or around virulence genes. This suggests a bias in favour of MNR regions within non-coding DNA, which suggests evolutionary pressure is being applied despite the fact that these areas of DNA are non-coding suggesting involvement in genomic regulation, i.e. if these non-coding regions are not under selective pressure and are random the probability of a string of repeated bases occurring would be small. For example, the probability of a 6bp poly-A repeat occurring would be 0.0002 and an 8bp poly-A repeat would be 0.00002, meaning these would be expected to appear randomly 5 and 0.5 times over the course of a 2.5Mbp genome respectively. Since they appear more frequently it suggests that when they appear, they have a function of some description.

A selection of non-coding DNA targets containing MNRs but still around virulence factors were selected to assess their suitability as profiling targets compared to non-coding regions that did not contain MNRs. The results from this were interesting, the target that split the isolate collection into the largest number of unique sequence types did contain MNRs but the second highest did not contain MNR repeats.

This shows that just because a section of non-coding DNA contains MNRs does not mean it will be more variable due to genome instability as suggested previously (74). However, one interesting aspect of the MNR containing non-coding regions was that 2/4 targets were found to contain various insertion sequences including GBSi1, IS1548 and ISSag8 and a *S. equi* transposase. This suggests that non-coding regions which contain MNR repeats are more likely to contain an insertion sequence which makes sense as 1) an insert into non-coding DNA would not inactivate an existing gene and 2) because MNR regions may either create an area of instability in the genome or may be complementary to the ends of insertion sequences.

The findings here disagree with the work of Diamant et al. (42) who proposed that MNR tracts within non-coding DNA are a source of hyper variability. They state that MNR containing non-coding DNA is more variable than housekeeping genes and the level of variation is independent of the variation of surrounding DNA. The data shown here suggest that MNR repeat regions are more likely to be variable if the genes surrounding them are highly variable. This was shown by the SAG0043-SAG0044 intragenic region which does not contain MNRs but is the most variable intragenic region and it is located near the highly variable targets established during the core genome work earlier. The levels of variation within these MNR regions is not significantly different to the housekeeping genes of the MLST scheme and is in most cases significantly lower than the core genes selected for profiling. Despite this, on average the diversity of the non-coding DNA with MNR repeats was higher than the diversity of the non-coding DNA that does not contain MNR repeats but not by enough to compensate for other factors such as the variation of DNA surrounding the non-coding repeats.

When looking at the sequences of these targets the variation that was shown was very rarely shown in the length of MNR repeat regions as opposed to the remainder of the sequence which is surprising. This may suggest an efficient DNA repair system or that any repeat regions are actively selected for which would make sense considering that MNR repeats are overrepresented in non-coding DNA which would suggest they are being actively selected for. Additionally, one reason for looking at MNR containing non-coding DNA regions was to identify targets that would be highly variable in a short stretch of sequence. This would allow pyrosequencing to be used, however analysis of the targets showed that in the approximately 150bp that could be generated by the biotage pyrosequencing platform available there would be insufficient variation to justify using these targets as sequence typing markers.

In conclusion, profiling schemes designed using non-coding DNA both with and without MNR repeats was unable to create a profiling system more discriminatory than the MLST scheme using fewer markers showing that the presence of MNR repeats in non-coding DNA alone is not sufficient to select profiling markers. Other work has created non-coding DNA based profiling systems by aligning corresponding non-coding DNA regions across a selection of genomes and selecting the most variable alignments (47) which would appear to be a better method for selecting profiling markers.

However 1) a region that is more variable does not necessarily represent the evolution of the species and has already been demonstrated to be inappropriate for long term evolutionary studies (136) and 2) if targets are going to be selected on the basis of variability of alignments, there is no reason to restrict this analysis to the non-coding genome, there may be coding regions that are more variable if that is what is desired.

Chapter 6

Combining Three Gene Profiling with MNR Profiling

6.0 Combining Three Gene Profiling and MNR profiling

6.1 Introduction

The aim of this project was to develop a two component profiling system that accurately reflects the phylogeny of GBS clinical isolates and gives indications into enhanced virulence. The first component of this was achieved by developing a three gene typing system which has been shown to be more discriminatory than the current MLST scheme. The second component looked at potential regulation of virulence genes by looking at MNR regions in coding and non-coding DNA. Although together these MNR containing regions did not produce a typing system to rival existing methods, individual targets were able to add discriminatory power and act as a potential virulence indicator for GBS.

6.2 Results

The results from profiling using non-coding regions containing MNRs were added to the three gene profiling method and the number of sequence types calculated (Figure 6.1). The SAK_1320-SAK_1319 (*scpB-lmb*) intragenic region which equalled the discriminatory power of *valS* and this region gave in combination with the 3 gene profiling markers resulted in 69 unique sequence types (Figure 6.1).

Figure 6.1: The number of unique sequence types per dataset

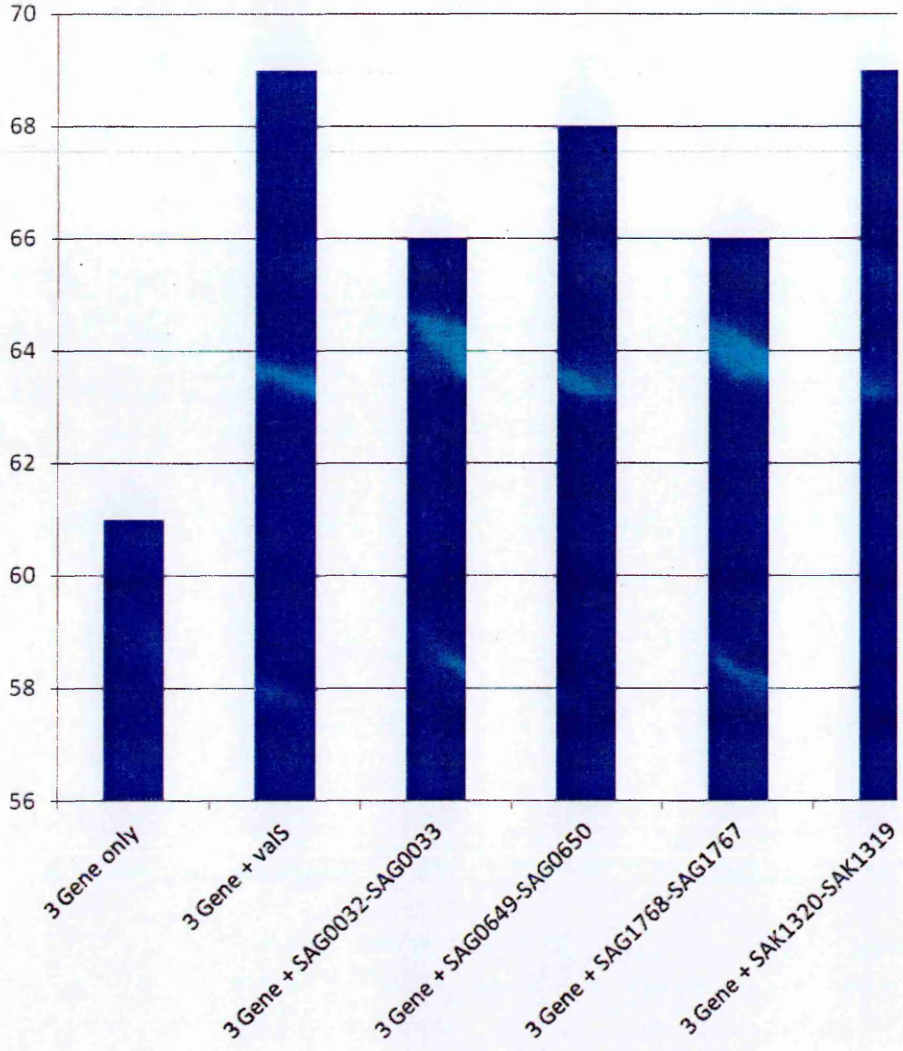
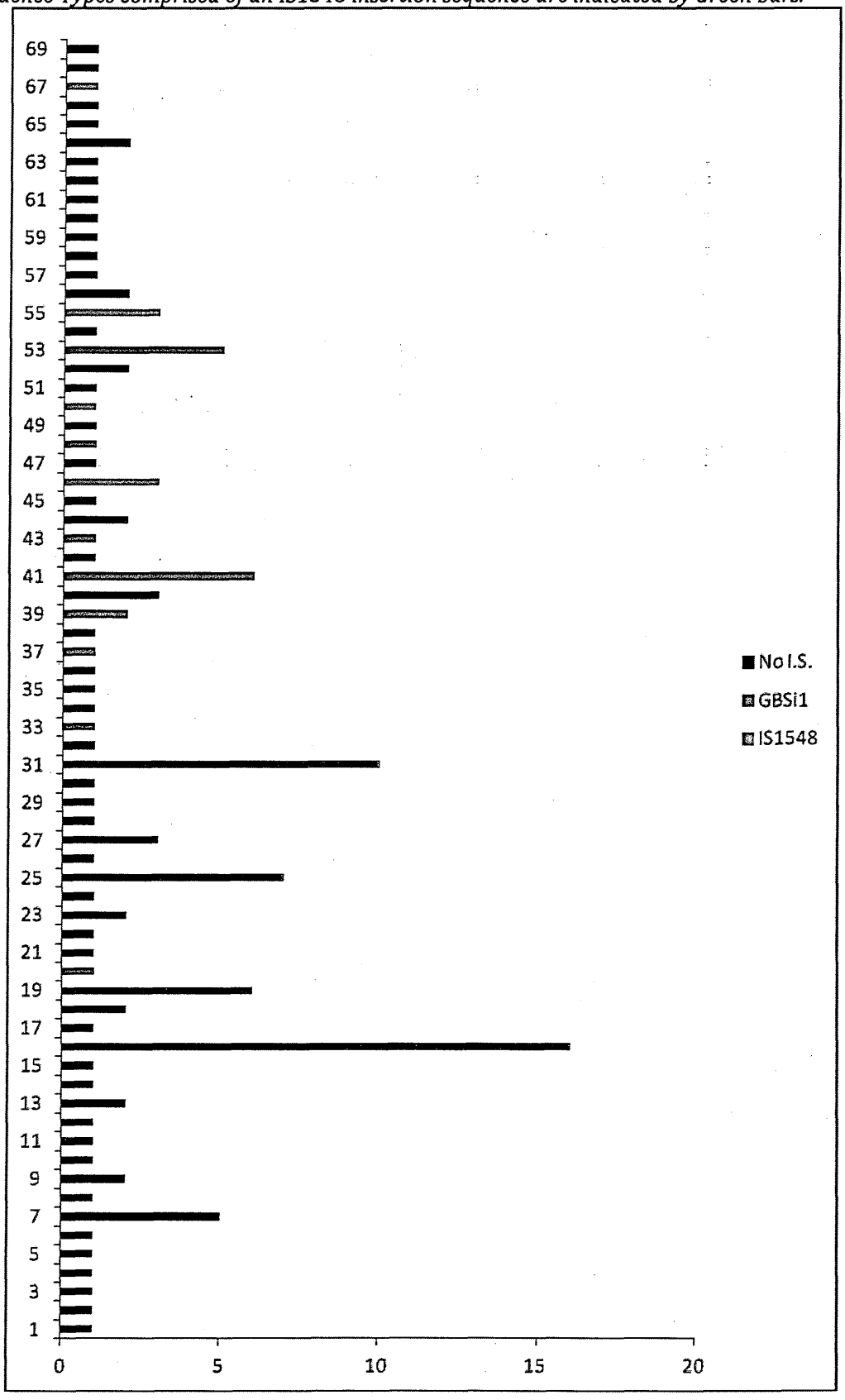


Figure 6.2: Distribution of Sequence Types using the Three Gene plus Insertion sequence typing scheme. Sequence Types comprised of a GBSi1 insertion sequence are indicated by Red bars and Sequence Types comprised of an IS1548 insertion sequence are indicated by Green bars.



These data show that the intragenic region between *scpB* and *lmb* does increase discriminatory power and results in more unique sequence types. However, two of the allele types found in the *scpB-lmb* intragenic region were the genomic inserts GBSi1 which were identified in 19 isolates and IS1548 which was identified in 7 isolates, giving a total of 26 (19.4%) sequenced isolates containing an insertion sequence.

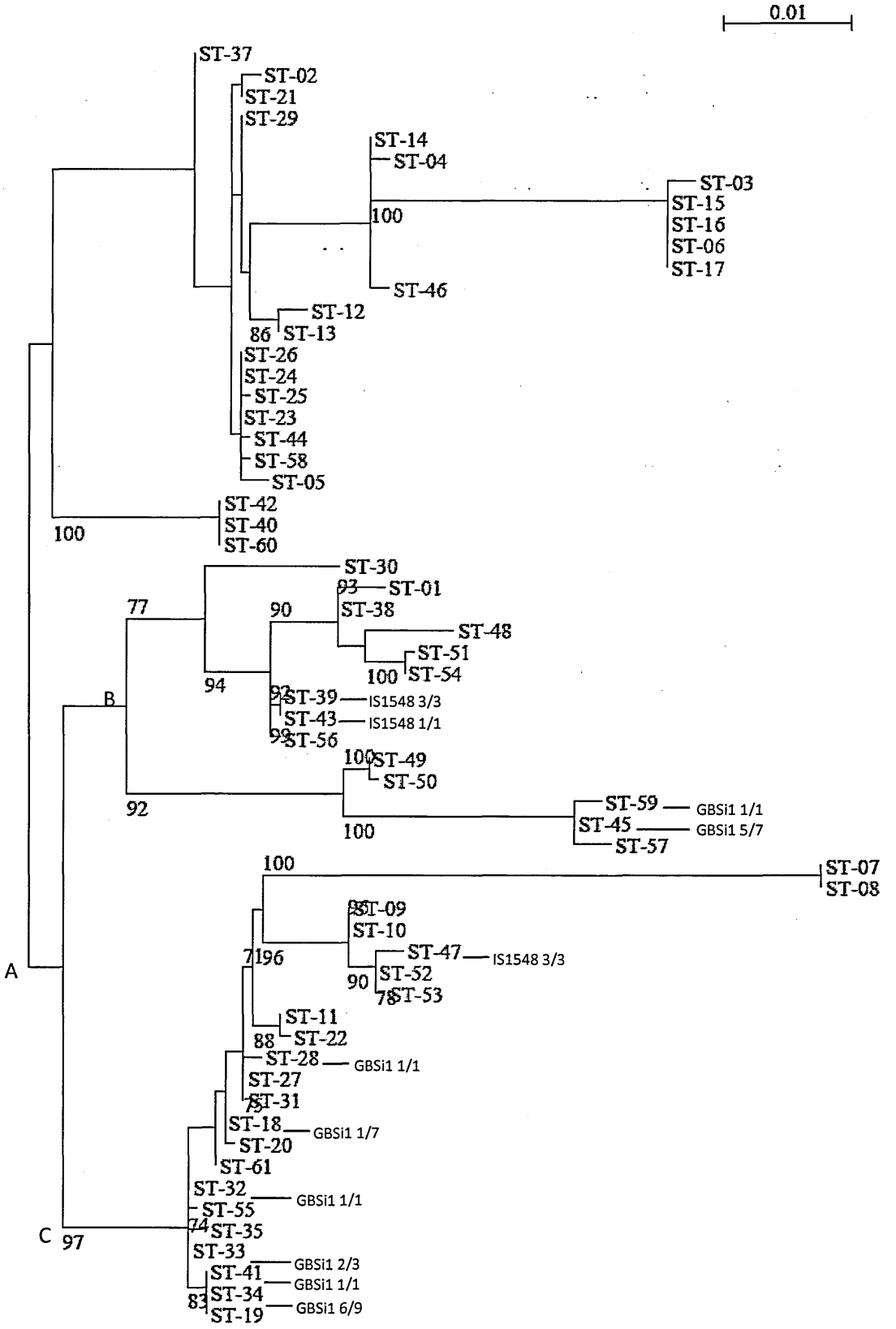
In total 12 three gene sequence types had isolates that contained inserts. Eight of these sequence types had an insert found in all strains of that type, of these the sequence types ST-28, ST-32, ST-36, ST-41 and ST-59 contained the GBSi1 insert and the sequence types ST-39, ST-43 and ST-47 contained the IS1548 insert. Four of the 3 gene sequence types had an insertion sequence in at least one isolate, ST-18 (7 isolates) had a GBSi1 insert in one isolate, ST-33 (3 isolates) had GBSi1 inserts in 2 isolates, ST-34 (9 isolates) had GBSi1 inserts in 6 isolates and ST-45 (7 isolates) had GBSi1 inserts in five isolates.

Using a loci with a relatively high proportion of insertion sequences means that this loci could not be used in phylogenetic analyses since any alignment would be misleading as each base is considered a separate evolutionary event which is not the case with an insertion of DNA sequence. Despite the limitations in using this marker in phylogenetic typing the presence of two distinct insertion sequences in a relatively high number of isolates presented a number of potential advantages. Firstly, figure 6.3 shows that these inserts were not found in the same 3 gene sequence type at this loci suggesting that insertion of either of these insertion sequences could be a marker of evolutionary lineage. Secondly, it has previously been shown that insertion sequences at this location may be partly responsible for increased virulence resulting in increased expression of *lmb*, a previously identified virulence factor involved in host cell invasion (2). Finally, the presence of this insert allowed further comparison of the MLST scheme and the 3 gene profiling method since isolates with insertion sequences would be expected to cluster together since, even considering GBS's open pan genome, an insertion is more likely to be a single event in the evolution of a subset of GBS strains. Therefore, mapping insertion sequences to the phylogenetic trees of the relationship between sequence types using the MLST scheme and the 3

gene profiling method should give an indication of how each scheme is representing the phylogeny of GBS strains.

When the sequence types containing inserts are added to the tree showing the relationships of the 3 gene sequence types (Figure 6.3), all isolates containing insertion sequences cluster within one clade (a section of the tree beginning with a common ancestor), clade A and within clade A there are two sub-clades (B and C). Clade B contains 2/3 of the sequence types (4/7 isolates) containing the IS1548 insertion sequence and 2/8 sequence types (6/19 isolates) containing the GBSi1 insertion sequence. Clade C contains 6/8 sequence types containing the GBSi1 insertion sequence (12/19 isolates) and 1/3 sequence types (3/7 isolates) containing the IS1548 insertion sequence.

Figure 6.3: The three gene profiling tree including the presence of inserts between the scpB-lmb genes and the number of isolates per ST with inserts. Points A, B and C indicate clades.



Finding that these insertion sequences were exclusive to a particular clade may suggest that these insertion sequences were gained around point A in the evolutionary history of GBS as indicated by

this tree. However, since these insertions were not present in all isolates after this point may suggest that the insertion sequence was either being lost because it was inherently unstable or, if this insertion sequence does increase virulence, because of selective pressure placed on the organism by treatment or invasive GBS.

When the positions of insertion sequences were mapped onto the MLST tree (Figure 6.4) it is clear that the insertion sequences were more spread throughout the tree when compared to the position of isolates on the 3 gene phylogenetic tree. Only 38.4% of the insertion sequences are found in the potential virulent clade identified earlier (Figure 4.8).

Figure 6.4: The MLST tree including the presence of inserts between the scpB-lmb genes

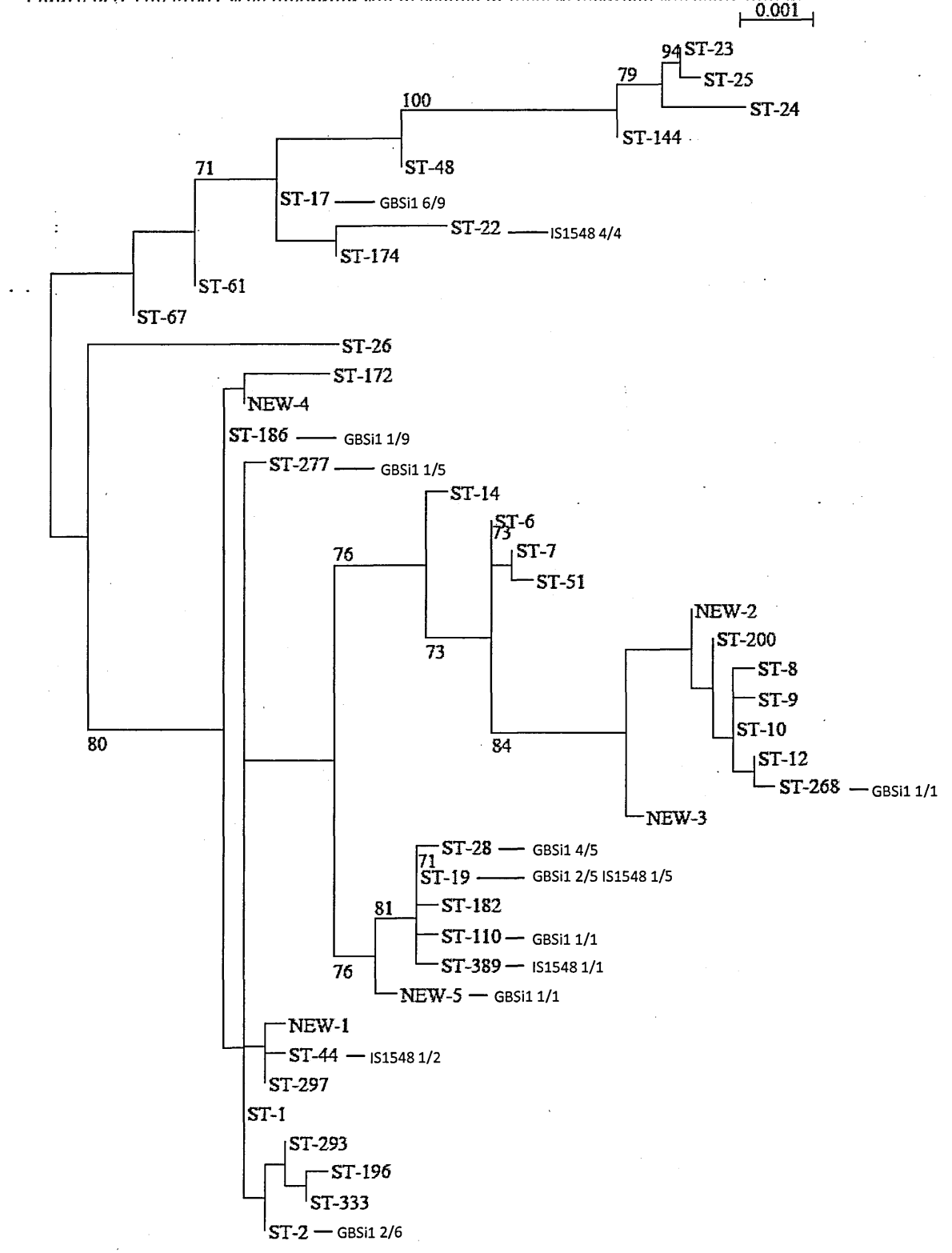


Figure 6.4 shows the diverse lineages which contain insertion sequences and also shows that unlike the three gene profiling method, one sequence type (ST-19) contains isolates containing both the GBSi1 and IS1548 insertion sequences which would not be expected if the sequence type is clonal suggesting that the MLST scheme is not accurately identifying phylogenetic lineages.

6.3 Discussion of Combining Three Gene Profiling and MNR Profiling

The aim of this section was to discover if the sequence typing using the three gene profiling system could be more discriminatory if supplemented with one of the markers selected from non-coding DNA. Since even though the MNR containing non-coding regions are not alone suitable for typing GBS, the addition of one of the MNR-targets was shown to improve the discriminatory power of the three gene profiling system and provided additional confirmation that this scheme is more phylogenetically representative that the existing MLST scheme.

The *scpB-lmb* intragenic region in 2/3 of the fully sequenced GBS genomes (A909 and NEM316) is a 235bp intragenic region containing MNR repeats, whereas in 2603V/R this region contains the gbsi1 insertion sequence. When combined to the three gene profiling method this region also produced the largest number of unique sequence types. The disadvantage of using this target is that due to the large proportion of insertion sequences this target cannot be used for sequence-based phylogenetic analysis. However, the use of this target is justified because the presence of IS1548 and GBSi1 at this location has been shown to increase the level of *lmb* which is a known virulence factor which binds to laminin and is therefore involved in GBS cells invading damaged epithelium (220). The two insertion sequences have been shown to increase the level of expression or *lmb* by variable degrees, expression is substantially increased when IS1548 is present and increased slightly in the presence of isolates containing the GBSi1 insertion sequence compared to isolates containing no insertion sequence (2). Due to this, the presence of an insertion sequence here can be considered a putative indicator of enhanced virulence, although the clinical impact is unknown and further work is required to determine this.

Including the presence and absence of insertions in analysis of the three gene sequence typing data confirms that the three gene profiling method is more phylogenetically accurate than the MLST scheme since the presence of any insertion sequence is limited to one clade in the 3 gene tree (Figure 6.3) and spread over the MLST tree (Figure 6.4) which would suggest that the three

gene profiling tree is hinting at a point in the evolution of the organism when it acquired these insertion sequences. Additionally, the MLST scheme contains multiple sequence types where the isolates within that sequence type contain different types of insertions, that is, a single sequence type can contain isolates that contain either the IS1548 or the GBSi1 insertion sequences at the same location. Since the presence of multiple different insertion sequences in sequence types that, according to the MLST scheme, are identical it is likely that the MLST scheme is not accurately modelling the relationships between these isolates. By contrast, the three gene profiling method never showed a sequence type that contained the multiple types of insertion sequences. However, it did show sequence types that have an insertion sequence in some but not all isolates, although the majority of sequence types are homogenous for insertions with seven out of the eleven sequence types that contain an insertion sequence having one in all isolates of that sequence type. The fact that this insertion sequence is so unstable at this point in the genome could suggest that streptococcal genomes are highly recombinant, as previously suggested by Springman et al. (221) or that insertions at this location are inherently unstable. The most likely explanation is that the three gene profiling method, while representing phylogeny better than the MLST scheme is still not perfect. This is understandable since in a typing scheme that perfectly represented the phylogeny it is highly unlikely that any two isolates would ever be placed into the same sequence type. This is because any two isolates gathered from different infections or even different sites of infection in the same patient are highly likely to show some difference at the genome level. A sequence typing system in which all isolates are guaranteed to fall into a different sequence type would be redundant as it would be increasingly difficult to identify common sources of infection without detailed analysis. The only way to generate a typing system that has 100% resolution would be to sequence the genomes. Even then, analyses of relationships are difficult. For example, with multiple sequenced genomes there are multiple methods for analysing the relationships such as using core genome SNPs in phylogenetic analysis (84) or cluster analysis based on genes that are not shared between all isolates (80), neither of which use all of the information generated from whole genome sequences.

In conclusion, this work set out to develop new profiling strategies based on different bioinformatic methods with varying degrees of success. Combining the elements of the highly discriminatory three gene profiling scheme with an element of the non-coding genome has the potential to become a profiling scheme that surpasses MLST in discriminatory power and accuracy. Combining a locus which may have real clinical implications will allow further study to elucidate the pathogenicity of GBS. This has the potential for future development of methods to determine if an isolate has the potential to cause disease.

Chapter 7

Discussion

7.0 Discussion

The aim of this project was to develop a two component sequence typing method that accurately reflects the evolution of GBS, is more discriminatory than the existing MLST scheme and could give an indication of the virulence of the organism. This was the goal because bacterial typing is one of the cornerstones of microbiology and has been performed by a large number of methods over many years. Each method has it's advantages and disadvantages and works differently when applied to different organisms. For example, the serotyping of GBS was developed by Rebecca Lancefield in 1934 (126) and is still used today to allow for historical comparisons and to allow vaccine development. Despite it's advantages, it does not discriminate very well and does not reflect the evolution of the organism. Another example is MLST, which despite being useful for studying the long term evolution of most organisms it is not as reliable in GBS as in other pathogens. This project aims to improve upon this scheme and develop a new sequence typing method that accurately reflects the evolution of GBS.

Bacterial typing is performed for a number of reasons. Firstly to differentiate strains of bacteria that can cause disease from those that cannot, for example only certain *E. coli* strains can cause disease and these can be differentiated by serotyping of the O, K, H and F antigens (14) which is in direct contrast to GBS where serotyping for epidemiology has shown that out of the 10 existing serotypes, serotypes Ia, III and V cause the majority of neonatal GBS infections in the West opposed to types VI and VIII in Japan (123), i.e. it appears that most if not all GBS serotype can cause disease but the prevalence of serotypes in a given locale is more important, that is, if serotypes VI and VIII were more prevalent in the UK then it is unlikely that the disease burden would be decreased, rather, the only difference is likely to be that infections show a different serotype. Also molecular typing methods have been used to track the spread of outbreaks such as Destro et al. who used RAPD and PFGE to trace the dissemination of *L. monocytogenes* in a shrimp processing plant (41) and is routinely used to confirm the source of existing outbreaks identified through studying the movements of and connections between patients.

Additionally, in research to answer key questions about the organism, for example, Sorensen et al. in 2010 (219) used MLST plus 8 additional housekeeping genes to shown that GBS clones are specifically adapted to either bovine or human hosts. However, sequence typing methods have been better able to study the population structure of other organisms than the GBS MLST scheme, the main sequence typing method for GBS. There are currently many commonly used methods for bacterial typing including serotyping, which uses antibodies against set proteins; phage typing which detects different susceptibilities to attack by certain bacteriophages and molecular methods such as PFGE, RAPD or MLST. Other methods are organism-specific and less common, for example in GBS three set genotyping (117), molecular serotyping (115) and a method referred to here as MLST+ which is MLST plus sequencing of other housekeeping schemes to add discriminatory power to MLST (219) all of which may be in response to limitations with the MLST scheme. The limitations of sequence typing are firstly lack of discrimination, that is, different bacteria that are different being placed into the same category and secondly the categories assigned do not represent the links between evolution of the organism as a whole. Hence, the goal of this project, to create a new molecular profiling system which is both more discriminatory and better represents evolution than the existing MLST scheme for GBS.

The approach used here was to develop bioinformatic pipelines for selecting genomic markers of phylogeny and variability in GBS and then test these in a laboratory setting. The methods developed here can also be applied to other microorganisms. The first method used was first developed by Konstantinidis et al. (119) and selects genomic targets based on their correlation to the average rate at which the genome is evolving and has previously been shown to generate a much more accurate phylogeny of *E. coli* and the *Salmonella*, *Burkholderia*, and *Shewanella* groups. Here this method is adapted for use at the strain level using an organism in which a large number of core genes show a high level of homology. The second is based on research that showed that MNR repeat regions are more highly variable that comparable DNA stretches (133) and may be involved in genomic regulation through slipped strand misspairing creating truncated translation products (165,242) and by other methods in non-coding sections of DNA (184). Also,

since it is known that the vast majority of virulence genes are shared between GBS sequenced isolates, work looking at sections of DNA that may be involved in regulation will focus on these genes as it may reveal interesting facts about GBS virulence.

Bacterial typing is important for diagnosis, treatment and epidemiological surveillance of infections. This is even more important in bacteria which exhibit high levels of antibiotic resistance, are involved in nosocomial or pandemic infections or for bacteria that are being studied for vaccine development. There are a variety of different typing methods which can broadly be split into four categories 1) serological. 2) DNA banding based methods. 3) DNA hybridization methods using nucleotide probes. 4) DNA sequencing based techniques. This section will discuss the general issues with typing and the advantages and disadvantages of each of the methods which have been applied to GBS previously and in this study.

In molecular biology, a bacterial species is defined as a homogeneous population. This can be determined experimentally through DNA hybridization which defines bacteria as being in the same species if they share at least 70% hybridization of DNA (249), by the homology between the 16S rRNA gene where an identity over 97% gene sequence similarity shows isolates to be in the same species (223) or by the Average Nucleotide Identity of the core genomes of a group of isolates where if the ANI is greater than 94% two genomes can be considered related (121).

Relatively recently there has been a large increase in the number of whole genome sequences available as a result of advances in whole genome sequencing technology and the corresponding decrease in cost. This has allowed comparative genomic methods to demonstrate that genetic diversity within a bacterial species was far greater than previously thought (15,65,132,231). This also introduced the concept of the bacterial pan-genome. This is every gene identified in a species subdivided into two components, the core-genome, which is comprised of genes shared by all sequenced strains in a species, and the accessory genome which is comprised of genes which are not found in all strains. Any two strains of the same species of bacteria may differ in gene content by as much as 30% and a species may show a much higher level of diversity (65,231). The concept

of an open or closed pan-genome has also been introduced. A closed pan-genome is a species with a relatively small pan-genome and sequencing of new strains will not add to the pan-genome. In contrast, an open pan-genome contains a large pan-genome and theoretically, sequencing additional strains of the species will result in the size of the pan-genome increasing indefinitely. For example, *Bacillus* genomes have been shown to have a closed pan-genome whereas GBS hypothetically has an open pan-genome (231). However, this may be due to sampling bias caused by using incomplete genome sequences. Genomic variations and typing can therefore be measured in two ways. Firstly variation can be measured by looking at core genes using either whole genome sequencing (83) or by using genes selected from the core genome to give a representation of genomic diversity such as in MLST (146). Also through more targeted methods which use genomic information to select more representative targets such as the typing schemes devised by Konstantinidis (118) and by extension of the work presented here. The advantage of methods which use the core genome is that it is possible to gain an idea of phylogeny and the evolutionary lineage of the species.

In contrast, it is also possible to type strains of bacteria using the pan-genome, for example Hiller el al. (88) used microarray analysis to determine the pan-genome of *S. pneumoniae* isolates and revealed that this method was more discriminatory than MLST which is also confirmed by Hall (79). However, using pan-genome microrrays for typing has a number of potential problems. Firstly, in the case of an open pan-genome it is possible than not all accessory genes will be identified in a microarray and secondly information regarding the phylogeny of the organism is limited if the core genome is not considered. Although it is possible that these problems will be overcome with the increasing use of whole genome sequencing allowing identification of new genes and also identifying the core genome of bacterial species so analysis of both core and accessory genomes can be performed in unison.

Despite the problems with pan-genome based typing, using the core genome also has its problems. The concept of pan-genome typing and the increased availability of whole genome sequencing challenges the use of the core genome for bacterial typing, by demonstrating that

typing methods using the core genome may not be able to evaluate the genetic diversity of a bacterial species. That is, despite the core genome or the representatives of it selected for strain typing such as MLST loci, may show a species to be identical but the strains may still have differences routed in its accessory genome. For example, the genomic diversity of GBS based on analysis of the whole genome was found to be inconsistent with MLST sequence types (159) and is also confirmed in this work.

This discrepancy may be explained by the fact that housekeeping genes used in MLST are selected from the core genome rather than the pan-genome. With a sequence based typing scheme using a limited number of genes it is sensible to use genes that will be found in all strains to obtain the maximum amount of information. In other words, one sequenced gene allows more potential discriminatory power than the presence or absence of one gene. Therefore, for pan-genome typing to be an effective typing tool would require either whole genome sequencing or microarray analysis which is costly and requires a longer turnaround time than traditional methods based on sequencing core genes. Typing based on specific genes from the pan-genome may, however, be particularly useful for typing based on genes that are involved with phenotypic or clinically important genomic factors.

Because intraspecies/genus diversity can result from SNPs, insertions or deletions, and recombination, it is sensible to consider the recombination rate of the selected bacteria when designing or selecting typing methods for any group of organisms. For example medically relevant pathogens such as species within the *Streptococcus* genus, the species *N. meningitidis*, the species *H. pylori*, and bacteria of the *Salmonella* species show high levels of genetic recombination (179), which needs to be considered. The traditional method to take recombination into account when developing sequence based typing schemes is to situate profiling loci around the genome so one recombination event will not disproportionally affect the overall sequence type as in the published MLST schemes and to use methods such as eBURST (59) which provides an analysis based on the presence and absence of particular allele types and is therefore ideal for analysis of recombination events between loci of sequence typing schemes. The only problem with this

approach is that allele types are not weighted on nucleotide identity so an allele type that is one nucleotide different from another is considered equally related to an allele type that shows significant differences at the nucleotide level. An alternative is to include additional profiling targets from the pan-genome to give an indication of the effects of recombination on the relationship between isolates, ideally ones that are clinically relevant.

Decoding the association between genotypes and phenotypic or epidemiological traits such as bacterial virulence, antibiotic resistance, host adaptation, geographic origin, pandemic, or epidemic outbreaks, is another major challenge of bacterial genotyping. Genotyping can be discriminatory depending on the loci selected for sequence typing. It is however, still important to take epidemiological traits into account when determining outbreaks of bacterial infections otherwise it is possible to falsely identify outbreaks. Although most typing markers are not genes directly responsible for virulence or antibiotic resistance, it is still possible to show correlations between specific sequence types and phenotypic or epidemiological traits by correlating phylogeny information with epidemiological data. For example, it is widely acknowledged that GBS sequence type 23 strains identified by MLST are highly virulent (141). Additionally, regardless of the typing method used, a larger collection of strains is also useful for linking genotypes with phenotypical and epidemiological traits (29).

Bacterial typing by serological methods is common in microbiology and is one of the oldest methods for bacterial typing, for example GBS is serotyped on the basis of ten type specific capsular polysaccharides (126,216) and is further subdivided using the localised protein antigens. This highlights the weakness of serotyping since using cell capsular polysaccharides GBS can only be subdivided into 10 distinct types and addition of the three cell protein markers increases the number of potential types to 30. This is a small number compared to the number of sequence types generated by both MLST (575 MLST sequence types) and the three gene profiling approach from this study (69 sequence types). However, this effect is species dependent. For example the *Salmonella enterica* species has over 2300 serovars split over six different subspecies (*enterica, salamae, arizonae, diarizonae, houtenae and indica)* whereas the MLST scheme contains 1525

sequence types according to the UCC MLST database. Another limitation of serotyping is that isolates can be non-typable, for example one study showed the number of non-typable GBS isolates was as high as 12% (116) and serotypes can also show cross reactivity resulting in a non-conclusive results.

Serotyping does have its advantages. Firstly, as the oldest typing method any isolates compared by this method can be studied relevant to the history of the organism, which is difficult with newer molecular profiling schemes, especially since an increase in genomic data is allowing more accurate sequence typing targets to be proposed potentially superseding established molecular profiling schemes. An additional advantage is that since serotyping markers are accessible to antibodies they may make good vaccine candidates. For example a recent meta-analysis has shown that the most common GBS serotypes are Ia, III and V and they account for 72% of infectious GBS isolates (95). However, if a trivalent vaccine was developed against these serotypes (175) it is likely the other GBS serotypes would increase in prevalence meaning any outbreaks would need to be monitored by serotyping to allow for further vaccine development. The GBS serotyping scheme is not the most discriminatory typing method available, as demonstrated in this work which showed that MLST, 3-gene typing and typing using non-coding DNA containing MNRs are all more discriminatory than serotyping, however, since serotyping it is essential for vaccine development of CPS based vaccines serotyping is unlikely to be stopped.

The DNA sequence is the primary genetic information of an organism and is used for differentiation and phylogenetic analysis of bacterial strains. DNA sequence-based genotyping methods are highly reproducible because they rely on unambiguous DNA sequences that are easily stored in online or local databases and are therefore easily compared between different laboratories without the inter-laboratory variation associated with DNA banding based methods. A number of DNA sequence databases exist, the three most common are GenBank, EMBL, and DDBJ the largest and most popular of these being NCBI's GenBank, which contains a massive amount of DNA sequences and is still increasing at an exponential rate. For example, in December 2000 there were $1x10^{10}$ base pairs and currently there is approximately $1.2x10^{11}$ bp of sequence

data. Sequence information is either in the form of complete genomes or locus-specific sequences from important targets for almost all known bacteria. Compared with other SNP genotyping methods, DNA sequencing is better suited for the identification of multiple SNPs within small regions of DNA, in addition to detecting sequence differences used to place bacteria into sequence types. The main advantage of DNA sequencing is that it allows the evaluation of the evolutionary forces that led to these differences, particularly when combined with phenotypic information such as antibiotic resistance. However, the usefulness of DNA sequencing largely depends on the targets selected and the desired function of the sequence typing scheme, for example 16S rDNA sequencing is ideal for identification at the genus level and at the species level but not suitable for differentiation between strains of the same bacteria. Whereas MLST cannot be used for identification of bacteria as it is species specific but is better than 16S rDNA sequencing at differentiating between isolates of the same species of bacteria.

In addition to these methods, there is MLST which is currently one of the most popular sequence typing methods for the characterization of bacterial strains and is currently the reference typing method for many bacteria. MLST typically involves the sequencing of 7 loci spread over the genome from genes present in all strains of a given bacteria. Because the loci are spread throughout the genome this gives the advantage that the results from one of the MLST schemes should not be heavily influenced by recombination events. That is, using a single locus could identify two isolates that on the whole are closely related as being massively different on the basis of one recombination event. Additionally, since loci are spread over the genome and have different functions they will not be evolving at the same rate as a genes located close to each other on the genome or are involved in the same function and therefore under the same selective pressures. As well as sharing the advantages of other sequence typing methods of portability and consistency between laboratories, MLST does however have its disadvantages. Firstly, analysis using eBURST which is a common method for MLST data analysis relies on alleles which are assigned to an arbitrary numbering system i.e. sequence types 1 and 2 are no more or less likely to be more related than sequence types 1 and 100, which is a weakness of this form of

phylogenetic analysis (32), of course other phylogenetic methods can be used but the high level of homology some MLST schemes show can give less reliable results. Secondly the use of conserved genes in MLST schemes often fails to detect the variability between strains which can be quite different at the genome level. Finally, sequencing seven loci is costly and time consuming when compared to other typing methods although the costs of sequencing have been decreasing for quite some time and are continuing to do so.

Most typing of GBS is now performed by MLST, partly because sequenced-based methods avoid problems of banding pattern based methods such as inter-laboratory variability and partly because of improved discrimination. MLST is now by far the most common GBS typing scheme, and as previously mentioned it has already been used in a number of studies on the population structure of in different countries including the UK (103,104), the USA (17), Sweden (141), Portugal (156), France (125), Israel (152), the Central African Republic and Senegal (23). As previously discussed the results from these studies suggest that invasive GBS is a worldwide clonal population which is highly unlikely. It is possible that using the new 3-gene typing system plus analysis of the SAK_1320 – SAK_1319 intragenic region will address this problem as it has been shown in this study that these targets represent the relationships between the 8 genome sequenced strains far better than MLST and that this new scheme is able to discriminate between isolates more effectively. However further work on a large and more varied strain collection is required to prove this hypothesis fully.

As an alternative to using coding regions for sequence typing various schemes using non-coding DNA have been developed, the first of these 16S–23S rRNA gene internal transcribed spacer (ITS) was the first non-coding sequence used for strain typing (66,199). Like the 16S rRNA gene it is common to all bacteria with the exception of *Rickettsiales* (138) and is usually present in multiple copies per genome (21). It is, however, much more variable than the 16S rRNA gene and is most commonly used for sub-typing bacteria. However, its variability and the fact it exists in multiple copies and given that the region varies in length between multiple copies in the genome, make direct sequencing difficult. Cloning based methods can overcome this problem but this makes the

process more difficult, time consuming and expensive. As well as single non-coding regions for typing, multiple regions can be selected for various organisms to create Multi Spacer Typing (MST). The assumption behind MST is that non-coding DNA is less affected by selective pressure than coding regions and therefore are more variable and will therefore provide more discriminatory strain typing systems (47). MST loci are selected either because of the presence of repeat regions which should also be more variable due to replication errors (74) or non-coding sequences which vary the most between different bacterial strains. This approach has proved useful for strain typing, for example MST based on six intergenic spacers divided 36 *Y. pestis* strains from three biovars from dental pulp of patients deceased from plague in the second and third pandemics into 19 sequence types (47). Additionally, MST has been applied to *Rickettsia conorii* (63), *Rickettsia prowazekii* (260), *Rickettsia sibirica* (62), *Coxiella burnetii* (69), *Bartonella henselae* (134,136), *Bartonella quintana* (61) and *Tropheryma whipplei* (137) and when MST schemes are compared to MLST schemes MST is shown to be more discriminatory (63,136). As with MLVA typing, it is possible that having incredibly variable loci for typing may make MST typing schemes too variable to provide useful evolutionary information. Additionally, since non-coding regions should not be affected by selective pressure (excluding non-coding DNA involved in regulation) it is unlikely that MST typing would give a true evolutionary history.

Multiple Loci VNTR Analysis (MLVA) is used to type organisms based on the number of tandem repeats found at any given locus that varies in copy number. These repeats and are dispersed widely in both human and bacterial genomes (144,240,245). In bacterial genomes, VNTR loci are found in non-coding regions as well as in genes and these non-coding VNTRs make good targets for strain typing because of their rapid evolution (63,170,240) and therefore MLVA tends to be more discriminatory than other methods.

Since MLVA, first used in 2000 has proven to be a highly discriminatory method for a number of bacteria it is now regarded as a reference typing method for many bacterial species, such as *Francisella tularensis* (57,99), *Bacillus anthracis* (89,111,140), *Yersinia pestis* (113), and *Mycobacterium tuberculosis* (129). Interestingly, these are usually considered to be highly

homologous groups that are difficult to type by other methods. MLVA has also been applied to other important human pathogens such as meticillin-resistant *Staphylococcus aureus* strains (230), *Burkholderia pseudomallei* (236), and *Clostridium difficile* (241). It was recently applied to GBS (186) and was able to split a strain collection into almost double the number of sequence types when compared to MLST.

However, the rapidly evolving nature of VNTR loci is also the main weakness of MLVA typing, it has been shown that VNTR loci may be too variable to provide reliable evolutionary information between closely related strains. For example MLVA typing of *Mycobacterium leprae*, has shown variation in the VNTR pattern between isolates of *M. leprae* biopsies from the same patient (162) showing that MLVA may be unsuitable for outbreak investigation and is more than likely unsuitable for long-term epidemiological surveillance (139). This effect however, may be species dependent as it has also been shown that MLVA for *Enterococcus faecium* is less discriminatory than PFGE and MLST (253). As a result the use of MVLA typing should be considered carefully for each organism.

Steps should also be undertaken to ensure than new MLVA schemes carefully select targets based on multiple sequenced genomes. Since previous research had shown that non-coding regions that contained MNR repeats were more variable than non-coding regions that do not contain them (42). This study however, showed that the presence of MNRs was not necessarily a predictor of enhanced variability. This could be species dependant because, for example GBS contains better DNA proofreading proteins to counter the effects of replication errors and therefore the results found in *E. coli* could not be translated to GBS. Alternatively, it is possible the MNR regions are not randomly occurring and they are maintained as they play a role in genomic regulation and are therefore under selective pressure to maintain relatively homogenous.

The point remains that typing based on the pan-genome will be more highly discriminatory and provide more accurate representation of phylogeny than any typing system that uses markers selected from the genome. However, it would be even more useful to sequence the whole

genome, giving scope for analysis of the core genome and the variable genome. In fact since the beginning of this project the technology for whole genome sequencing has improved dramatically and the cost has decreased significantly even to the point that one of the most frequent questions about this project is "won't this be irrelevant soon because we will just sequence the genomes of everything?" To answer that, the short answer is yes, but the long answer is not for a while. As Amara's law states, "We tend to overestimate the effect of a technology in the short run and underestimate the effect in the long run."

Whole genome sequencing is already being used to study the evolution of pathogens for example, Harris et al. (84) used the illumina genome analyser to generate sequence data for 63 MRSA isolates. Further work to elucidate the population structure of GBS should use these methods and take advantage of new technologies to sequence a large number of GBS genomes to reveal the actual population structure of GBS opposed to the current view from the MLST typing system that the worlds GBS population is essentially similar. A number of different methods of analysing the data could be employed including looking at the core genome to determine the evolutionary history of the organism (83) or typing based on the pan-genome proposed by Hall et al. who used a microarray approach to show that pan genome typing is able to differentiate the genome sequenced strains (79). This research is possible with current technologies and could be performed now especially considering the large number of GBS genomes being published by the JCVI recently meaning population structure could be studied by any research group without the associated costs of sequencing a large number of genomes.

The holy grail of next generation sequencing in microbiology would be the ability to sequence multiple bacterial genomes and deliver and interpret the resultant sequence information in "real-time" in order to use sequence information for patient management or for high resolution outbreak tracking. As recently as May 2012 there was no examples of this occurring (49) but most recent reviews agree that the rapid advances in whole genome sequencing will become more and more prevalent in clinical microbiology. The advantages of this being that three key areas of microbiology can be addressed in a single workflow thereby replacing many current complex and

expensive techniques, firstly, species identification at a much higher resolution than current methods, secondly, an ability to test an organisms properties such as virulence and resistance to antibiotics and finally, perhaps most usefully, to monitor the spread and emergence of human pathogens (43). Additional advantages to using next generation sequencing technology includes clinical metagenomics, a weakness of Sanger sequencing since firstly without expensive and time consuming random amplification and cloning techniques (22) only the most prevalent of any selected target sequence will be represented compared to NGS platforms that will give all manner of information of quasi-species and multiple infections (172).

Despite this technology being difficult to bring into routine clinical microbiology at this time, it is already being used frequently for research purposes. For example a recent study by Claudio published in June 2012 has used the Illumina MiSeq to sequence 14 MRSA genomes, 7 of which were associated with a hospital outbreak and 7 which were associated with carriage of MRSA or bacteremia in the same hospital which showed a previously missed transmission event between two patients with bacteremia who were not part of the outbreak. However this study did highlight difficulties in analysis, epidemiological analyses suggested all seven outbreak strains were linked however, sequence analysis showed that one of the outbreak isolates was much more distantly related than the other outbreak isolates. This either highlights the difficulty of imposing a simple identity based threshold to classify isolates as belonging to the same outbreak or it demonstrates that current epidemiological methods can incorrectly place unrelated isolates together. From a patient management point of view genomic data also allowed the creation of an artificial "resistome" of antibiotic resistance genes which would allow intelligent selection of antibiotics (33). However, costs using this approach are still high with each genome sequence costing approximately £285 compared to around £12 for sequence typing. Additionally this study was a retroactive analysis, whilst the timeframe the study was conducted in, means this method could theoretically be applied in a real clinical setting it was not in this case meaning that there is still research to be done and a detailed consideration of costs and benefits of using such an approach before this type of analysis would be applied routinely in a clinical setting.

That being said current advances in whole genome sequencing technology will see sequence typing replaced in the near future by diagnostic whole genome sequencing which would provide much more information about the genomic aspects of infectious disease and could lead to giant leaps in the understanding but these methods are currently not available or cost efficient to most laboratories. Another technology that could provide cost efficient real time genomic information is the Life technologies Ion PGM (personal genome machine), which in my view is the third generation sequencing technology most suited to be used as a solution for microbial whole genome sequencing due to its relatively low cost and proven ability to rapidly generate the genome sequence of the recent Enterohemorrhagic *E. coli* O104:H4 isolate responsible for the recent outbreak in Germany (160). Despite the potential with this platform, the cost of a GBS genome would be £316 using a single 10Mb 314 chip which would theoretically give approximately 5x coverage of a GBS genome, but is still a little on the low side practically since the above sequencing required the use of ten 314 chips. However, using the 316 chip which were not available when this study was carried out, have a capacity of 100MBp so the cost of sequencing a GBS genome would be £700 and give approximately 50x coverage (Miles Collier, Life Technologies, personal correspondence). This would be even cheaper if a number of samples could be multiplexed, for example for ~10x coverage using a 316 chip five samples can be processed per chip and would cost £140 per genome and there is no reason to believe this cost will not fall further since the sequencing capacity of this platform essentially relies on how many semiconductors can be placed onto a chip and since Moore's law states that the number of transistors that can be placed inexpensively on an integrated circuit doubles approximately every two years the capacity of the technology can only increase, decreasing the costs per base.

Although the advantages of whole genome sequencing in clinical microbiology would be innumerable several issues must be resolved before whole genome sequencing becomes routine in public health. The first of these issues being cost, as discussed above the latest research places the cost per genome at £285 per genome and a theoretical lower end of £140 which are both higher than the £12 that sequence typing using four targets would cost. Of course there is no

doubt that the costs of sequencing of whole genomes has fallen dramatically and will continue to fall, the only question is which technology will be the first to drop costs enough to make whole genome sequencing viable on a routine basis.

Which technology will be the main driver of whole genome sequencing in clinical microbiology is still a matter of much debate and from the point of view of the clinician or the technician what technology is used is unimportant. However, it is still a question worth considering. A recent review by Quail et al. (185) concludes that the Ion Torrent, Pacific Biosciences and Illumina MiSeq sequencers all able to generate usable sequence but there are differences in quality of the data generated and supported applications. For example, the error rates for the Illumina platforms and the Ion Torrent were 0.4% and 1.78% respectively and a rather high 13% for the Pacific Biosciences platform. This resulted in the Illumina platform producing 76.45% of reads without a single mismatch or indel opposed to 15.92% and 0% error free reads of the Ion Torrent and Pacific Biosciences sequencing platforms respectively. These figures taken together should indicate the Illumina platform will be the main driver of whole genome sequencing technology. Quail et al. (184) found that as long as there was sufficient coverage the errors do not make a significant impact on the final analysed sequence. In fact, they conclude for sequencing of microbial genomes that either the Ion Torrent or the Illumina platform will work well. They do however conclude that the Pacific Biosciences platform is currently unsuitable. Although it is worth noting that the Illumina chemistry is a tried and tested formula which although being a field leader now did have initial difficulties and the Ion Torrent and Pacific Biosciences platforms may well be at that stage of development now and who knows how they will perform in just a few years.

Additional problems with whole genome sequence data in a clinical microbiology setting include, there being a lack of staff with the knowledge or experience to deal with large volumes of genomic information, although this may be partially overcome with improvements in computer technology and software to make large scale analysis more user friendly and faster. It also highlights the importance of training staff to carry out bioinformatic analyses as well as the importance of the techniques used in this study since the majority of work in the evolution of

pathogens using whole genome sequence will either use the core genome to perform phylogenetic analysis on or the pan-genome to analyse the presence or absence of certain genes.

Finally, data storage is also an issue, for example the Sanger centre has previously had to delete experimental data before any meaningful analysis has taken place due to space restrictions (Jose Afuso Guerra Assuncao, personal correspondence), although, again, this problem will become less of an issue as either technology improves and storage becomes cheaper or scientific centres move more towards central data storage and cloud computing. Although there is no reason to believe that any of these problems are insurmountable sequence typing based on selected loci will be used in clinical microbiology for quite a while.

Therefore, as long as sequence typing is based on sequencing of selected loci, it makes sense to select those loci sensibly. This project explored different methods for selecting targets for molecular typing systems and from this I showed that it is possible to select markers that are more accurate and discriminatory by employing bioinformatics methods. The best of these methods being an adaptation of the Konstantinidis method (119) and selecting targets based on results from this. However using a bioinformatic approach alone is not always sufficient as I demonstrated and any analysis performed in-silico must be confirmed in the laboratory. In this study this was demonstrated by showing that performing an analysis and selecting the top 3 targets is not guaranteed to produce the most effective typing system. It is however an excellent starting point but selected targets must be verified experimentally by sequencing a collection of isolates and the results compared to an existing gold standard method or if no existing method exists by comparing multiple combinations of targets to select the best combination. At the time of the study there were only a limited number of sequenced genomes and information revealed from these may be applicable in varying degrees to real world isolates however since more sequenced genomes are now available the target selection process would be improved. Despite that, this project has demonstrated that fully investigating the sequence typing methods, rather than simply picking your favourite seven housekeeping genes is worth the effort due to increased accuracy in the data generated, decreased cost and time saved.

7.1 Future Work

Despite what is shown about the benefits of intelligent selection of sequence typing targets in this project further work is required to investigate the population structure of GBS and in my opinion the best way to do this would be to sequence the genomes of as many GBS isolates as feasible. However since this is not yet practical the typing system developed here should be applied to as wide a collection of clinical isolates as possible. As mentioned earlier the strain collection used in this study was a collection of UK clinical isolates, a more varied strain collection would enable a new set of questions to be answered and gain more reliable insight into the population structure of GBS. Specifically a collection of isolates from outside the UK would show the ability of this scheme to differentiate strains on the basis of country of origin. Other questions that could be answered include attempting to differentiate between invasive and non-invasive isolates in order to identify infectious vs. non-infectious lineages and would require sequence typing of non-invasive isolates which are difficult to obtain. That is, to obtain a sample that is a truly non-infectious isolate an expectant mother would have to be screened for GBS, proven positive and then refuse or be unable to receive intrapartum antibiotics to prove that infection was not passed onto the neonate. But it is also worth noting that not enough is known about host factors to estimate if a particular strain would cause infection in a different host. An alternative to this would be to use bovine or piscine isolates as a substitute for non-invasive isolates (55) but it is again uncertain if these could infect neonates. A recent study has suggested that zoonotic infection is possible by comparing colonisation of 68 families with their livestock (151) although this study did use a number of typing methods it is possible that lack of discrimination is affecting these results.

It would also be valuable to compare all existing GBS typing methods on the same isolates at the same time. There are already for example, publications that use MLST and PFGE (156) or MLST and serotyping (16) but there is no single comparison of all typing methods and since typing of organisms is one of the cornerstones of microbiology it would seem worthwhile comparing all typing methods on a selection of isolates isolated from as many different sources as possible to

definitively determine the best typing system, or rather, to know what system gives you what information and therefore when each method should be applied. Perhaps through collaboration with other national centres to get a generic scheme.

Further work into the affects of MNR repeats on genomic regulation should focus on the *cylA* and *cylB* genes since there are already a number of genomic controls identified for expression of capsule genes (114). This work could proceed in one of two ways. Firstly using a method like Fluorescence-activated cell sorting (FACS) to separate cells expressing the cylA and cylB proteins from those expressing one or neither followed by sequencing the section of the gene containing the MNR repeat region to determine if there is a change in the length of the MNR repeat region which would be the most likely cause of a lack of expression. Alternatively, it would be possible to try to force phase variation by growing liquid cultures and deep sequencing loci containing MNR repeats using 454 amplicon sequencing to determine the proportion of sequences with a frame shift mutation since it is already known that 5% of GBS clinical isolates do not show haemolytic activity (183).

Further work should also be carried out with other streptococcus species. MLST results for *S. pneumoniae* and *S. pyogenes* also seem to have a large proportion of sequence types that are widely distributed globally making it difficult to assume that the sequence typing performed on these organisms is any more reliable than that of GBS. This would suggest that applying this target selection approach on these organisms either at the species level or with all members of the streptococcus group could allow study of the population structure of these medically relevant pathogens. Advantages of this approach would be the additional information generated from so many extra genomes and a stronger evolutionary context to any targets selected and the original Konstantinidis method has already been shown to be adept at selecting targets at the genus level before the adaptations in this study. The disadvantages of using such a high number of genomes is that performing this method on all sequenced streptococcus genomes of which there are nearly 1,000 in the NCBI genome database may limit discrimination between strains at the species level, computationally this analysis would be very expensive and finally the number of core genes may

be reduced to a number that limits target selection. It was suggested by Tettelin et al. the number of core genes will decrease indefinitely as newly sequenced genomes are added (231) and so the effects of adding in a large number of general streptococcus genomes could have a big impact. Other research has however suggested a pan streptococcus typing system is feasible since it has already been shown that 26 streptococcus genomes have a core genome that plateaus at around 600 genes (132) although further analysis would be required to ensure this number would not fall with the addition of so many additional genomes.

Finally, before whole genome sequencing technology is ready to be applied routinely, any new sequence typing methods being developed for different organisms should consider using genomic information to make informed decisions about target selection. This has been successfully applied to GBS in this study and could easily be applied to other organisms.

Chapter 8

References

8.0 References

1. **Ahmod, N.** 2010. King's College, University of London. Elucidating the Complex Phylogeny of the Genus *Bacillus.*

2. **Al Safadi, R., S. Amor, G. Hery-Arnaud, B. Spellerberg, P. Lanotte, L. Mereghetti et al.** 2010. Enhanced expression of *lmb* gene encoding laminin-binding protein in *Streptococcus agalactiae* strains harboring IS1548 in scpB-lmb intergenic region. PLoS.ONE. **5**:e10794.

3. **Alm, R. A., L. S. Ling, D. T. Moir, B. L. King, E. D. Brown, P. C. Doig et al.** 1999. Genomic-sequence comparison of two unrelated isolates of the human gastric pathogen *Helicobacter pylori.* Nature **397**:176-180.

4. **Angata, T. and A. Varki.** 2002. Chemical diversity in the sialic acids and related alpha-keto acids: an evolutionary perspective. Chem.Rev. **102**:439-469.

5. **Angel, C. S., M. Ruzek, and M. K. Hostetter.** 1994. Degradation of C3 by *Streptococcus pneumoniae.* J.Infect.Dis. **170**:600-608.

6. **Areschoug, T., M. Stalhammar-Carlemalm, I. Karlsson, and G. Lindahl.** 2002. Streptococcal beta protein has separate binding sites for human factor H and IgA-Fc. J.Biol.Chem. **277**:12642-12648.

7. **Azzari, C., M. Moriondo, G. Indolfi, C. Massai, L. Becciolini, M. M. de et al.** 2008. Molecular detection methods and serotyping performed directly on clinical samples improve diagnostic sensitivity and reveal increased incidence of invasive disease by *Streptococcus pneumoniae* in Italian children. J.Med.Microbiol. **57**:1205-1212.

8. **Baker, C. J.** 1996. Inadequacy of rapid immunoassays for intrapartum detection of group B streptococcal carriers. Obstet.Gynecol. **88**:51-55.

9. **Baker, C. J. and D. L. Kasper.** 1976. Correlation of maternal antibody deficiency with susceptibility to neonatal group B streptococcal infection. N.Engl.J.Med. **294**:753-756.

10. **Balter, S., E. R. Zell, K. L. O'Brien, A. Roome, H. Noga, M. Thayu et al.** 2003. Impact of intrapartum antibiotics on the care and evaluation of the neonate. Pediatr.Infect.Dis.J. **22**:853-857.

11. **Bateman, A., E. Birney, R. Durbin, S. R. Eddy, K. L. Howe, and E. L. Sonnhammer.** 2000. The Pfam protein families database. Nucleic Acids Res. **28**:263-266.

12. **Bayliss, J., R. Moser, S. Bowden, and C. A. McLean.** 2010. Characterisation of single nucleotide polymorphisms in the genome of JC polyomavirus using MALDI TOF mass spectrometry. J.Virol.Methods **164**:63-67.

13. **Bedford, H., L. J. de, S. Halket, C. Peckham, R. Hurley, and D. Harvey.** 2001. Meningitis in infancy in England and Wales: follow up at age 5 years. BMJ **323**:533-536.

14. **Bettelheim, K. A.** 2003. Non-O157 verotoxin-producing *Escherichia coli*: a problem, paradox, and paradigm. Exp.Biol.Med.(Maywood.) **228**:333-344.

15. **Binnewies, T. T., Y. Motro, P. F. Hallin, O. Lund, D. Dunn, T. La et al.** 2006. Ten years of bacterial genome sequencing: comparative-genomics-based discoveries. Funct.Integr.Genomics **6**:165-185.

16. **Bisharat, N., D. W. Crook, J. Leigh, R. M. Harding, P. N. Ward, T. J. Coffey et al.** 2004. Hyperinvasive neonatal group B streptococcus has arisen from a bovine ancestor. J.Clin.Microbiol. **42**:2161-2167.

17. **Bohnsack, J. F., A. Whiting, M. Gottschalk, D. M. Dunn, R. Weiss, P. H. Azimi et al.** 2008. The Population Structure of Invasive and Colonizing Strains of *Streptococcus agalactiae* from Neonates of Six U.S. Academic Centers from 1995 to 1999. J.Clin.Microbiol.

18. **Bohnsack, J. F., A. A. Whiting, R. D. Bradford, B. K. Van Frank, S. Takahashi, and E. E. Adderson.** 2002. Long-range mapping of the *Streptococcus agalactiae* phylogenetic lineage restriction digest pattern type III-3 reveals clustering of virulence genes. Infect.Immun. **70**:134-139.

19. **Bolduc, G. R., M. J. Baron, C. Gravekamp, C. S. Lachenauer, and L. C. Madoff.** 2002. The alpha C protein mediates internalization of group B Streptococcus within human cervical epithelial cells. Cell Microbiol. **4**:751-758.

20. **Borchardt, S. M., B. Foxman, D. O. Chaffin, C. E. Rubens, P. A. Tallman, S. D. Manning et al.** 2004. Comparison of DNA dot blot hybridization and lancefield capillary precipitin methods for group B streptococcal capsular typing. J.Clin.Microbiol. **42**:146-150.

21. **Boyer, S. L., V. R. Flechtner, and J. R. Johansen.** 2001. Is the 16S-23S rRNA internal transcribed spacer region a good tool for use in molecular systematics and population genetics? A case study in *cyanobacteria*. Mol.Biol.Evol. **18**:1057-1069.

22. **Braham, S., M. Iturriza-Gomara, and J. Gray.** 2009. Optimisation of a single-primer sequence-independent amplification (SP-SIA) assay: detection of previously undetectable norovirus strains associated with outbreaks of gastroenteritis. J.Virol.Methods **158**:30-34.

23. **Brochet, M., E. Couve, R. Bercion, J. M. Sire, and P. Glaser.** 2008. Population structure of human isolates of *Streptococcus agalactiae* from Dakar and Bangui. J.Clin.Microbiol.

24. **Brodeur, B. R., M. Boyer, I. Charlebois, J. Hamel, F. Couture, C. R. Rioux et al.** 2000. Identification of group B streptococcal Sip protein, which elicits cross-protective immunity. Infect.Immun. **68**:5610-5618.

25. **Buhimschi, C. S., I. A. Buhimschi, S. bdel-Razeq, V. A. Rosenberg, S. F. Thung, G. Zhao et al.** 2007. Proteomic biomarkers of intra-amniotic inflammation: relationship with funisitis and early-onset sepsis in the premature neonate. Pediatr.Res. **61**:318-324.

26. **Buhimschi, C. S., A. T. Dulay, S. bdel-Razeq, G. Zhao, S. Lee, E. J. Hodgson et al.** 2009. Fetal inflammatory response in women with proteomic biomarkers characteristic of intra-amniotic inflammation and preterm birth. BJOG. **116**:257-267.

27. **Cai, Y., F. Kong, and G. L. Gilbert.** 2007. Three new macrolide efflux (*mef*) gene variants in *Streptococcus agalactiae*. J.Clin.Microbiol. **45**:2754-2755.

28. **Carver, T. J., K. M. Rutherford, M. Berriman, M. A. Rajandream, B. G. Barrell, and J. Parkhill.** 2005. ACT: the Artemis Comparison Tool. Bioinformatics. **21**:3422-3423.

29. **Caws, M., G. Thwaites, S. Dunstan, T. R. Hawn, N. Thi Ngoc Lan, N. T. T. Thuong et al.** 2008. The Influence of Host and Bacterial Genotype on the Development of Disseminated Disease with *Mycobacterium tuberculosis*. PLoS Pathog **4**:e1000034.

30. **Chaffin, D. O., S. B. Beres, H. H. Yim, and C. E. Rubens.** 2000. The serotype of type Ia and III group B streptococci is determined by the polymerase gene within the polycistronic capsule operon. J.Bacteriol. **182**:4466-4477.

31. **Chen, L., J. Yang, J. Yu, Z. Yao, L. Sun, Y. Shen et al.** 2005. VFDB: a reference database for bacterial virulence factors. Nucleic Acids Res. **33**:D325-D328.

32. **Clarke, S. C.** 2002. Nucleotide sequence-based typing of bacteria and the impact of automation. BioEssays **24**:858-862.

33. **Claudio, ., M. T. G. Holden, M. J. Ellington, E. J. P. Cartwright, N. M. Brown et al.** 2012. Rapid Whole-Genome Sequencing for Investigation of a Neonatal MRSA Outbreak. N Engl J Med **366**:2267-2275.

34. **Clawson, M. L., J. E. Keen, T. P. Smith, L. M. Durso, T. G. McDaneld, R. E. Mandrell et al.** 2009. Phylogenetic classification of *Escherichia coli* O157:H7 strains of human and bovine origin using a novel set of nucleotide polymorphisms. Genome Biol. **10**:R56.

35. **Cleary, P. P., U. Prahbu, J. B. Dale, D. E. Wexler, and J. Handley.** 1992. Streptococcal C5a peptidase is a highly specific endopeptidase. Infect.Immun. **60**:5219-5223.

36. **Cousens, S., H. Blencowe, M. Gravett, and J. E. Lawn.** 2010. Antibiotics for pre-term pre-labour rupture of membranes: prevention of neonatal deaths due to complications of pre-term birth and infection. Int.J.Epidemiol. **39 Suppl 1**:i134-i143.

37. **Coutte, L., S. Alonso, N. Reveneau, E. Willery, B. Quatannens, C. Locht et al.** 2003. Role of adhesin release for mucosal colonization by a bacterial pathogen. J.Exp.Med. **197**:735-742.

38. **Cromwell, D., T. Joffe, R. Hughes, D. Murphy, C. Dhillon, and M. J. van der.** 2008. The local adaptation of national recommendations for preventing early-onset neonatal Group B Streptococcal disease in UK maternity units. J.Health Serv.Res.Policy **13 Suppl 2**:52-57.

39. **Dela Cruz, W. P., J. Y. Richardson, J. M. Broestler, J. A. Thornton, and P. J. Danaher.** 2007. Rapid determination of macrolide and lincosamide resistance in group B streptococcus isolated from vaginal-rectal swabs. Infect.Dis.Obstet.Gynecol. **2007**:46581.

40. **Delcher, A. L., D. Harmon, S. Kasif, O. White, and S. L. Salzberg.** 1999. Improved microbial gene identification with GLIMMER. Nucleic Acids Res. **27**:4636-4641.

41. **Destro, M. T., M. Leitao, and J. M. Farber.** 1996. Use of Molecular Typing Methods To Trace the Dissemination of *Listeria monocytogenes* in a Shrimp Processing Plant. Appl.Environ.Microbiol. **62**:1852-1853.

42. **Diamant, E., Y. Palti, R. Gur-Arie, H. Cohen, E. M. Hallerman, and Y. Kashi.** 2004. Phylogeny and strain typing of *Escherichia coli*, inferred from variation at mononucleotide repeat loci. Appl.Environ.Microbiol. **70**:2464-2473.

43. **Didelot, X., R. Bowden, D. J. Wilson, T. E. Peto, and D. W. Crook.** 2012. Transforming clinical microbiology with bacterial genome sequencing. Nat Rev.Genet **13**:601-612.

44. **Diekema, D. J., J. I. Andrews, H. Huynh, P. R. Rhomberg, S. R. Doktor, J. Beyer et al.** 2003. Molecular epidemiology of macrolide resistance in neonatal bloodstream isolates of group B streptococci. J.Clin.Microbiol. **41**:2659-2661.

45. **Doran, K. S., V. M. Benoit, R. E. Gertz, B. Beall, and V. Nizet.** 2002. Late-onset group B streptococcal infection in identical twins: insight to disease pathogenesis. J.Perinatol. **22**:326-330.

46. **Doran, K. S. and V. Nizet.** 2004. Molecular pathogenesis of neonatal group B streptococcal infection: no longer in its infancy. Mol.Microbiol. **54**:23-31.

47. **Drancourt, M., V. Roux, L. V. Dang, L. Tran-Hung, D. Castex, V. Chenal-Francisque et al.** 2004. Genotyping, Orientalis-like *Yersinia pestis*, and plague pandemics. Emerg.Infect.Dis. **10**:1585-1592.

48. **Duarte, R. S., O. P. Miranda, B. C. Bellei, M. A. Brito, and L. M. Teixeira.** 2004. Phenotypic and molecular characteristics of *Streptococcus agalactiae* isolates recovered from milk of dairy cows in Brazil. J.Clin.Microbiol. **42**:4214-4222.

49. **Dunne, W. M., Jr., L. F. Westblade, and B. Ford.** 2012. Next-generation and whole-genome sequencing in the diagnostic clinical microbiology laboratory. Eur.J.Clin.Microbiol.Infect.Dis. **31**:1719-1726.

50. **Durso, L. M., J. L. Bono, and J. E. Keen.** 2005. Molecular serotyping of *Escherichia coli* O26:H11. Appl.Environ.Microbiol. **71**:4941-4944.

51. **Edmond, K. and A. Zaidi.** 2010. New approaches to preventing, diagnosing, and treating neonatal sepsis. PLoS.Med. **7**:e1000213.

52. **El Aila, N. A., I. Tency, G. Claeys, B. Saerens, B. E. De, M. Temmerman et al.** 2009. Genotyping of *Streptococcus agalactiae* (group B streptococci) isolated from vaginal and rectal swabs of women at 35-37 weeks of pregnancy. BMC.Infect.Dis. **9**:153.

53. **El Aila, N. A., I. Tency, G. Claeys, H. Verstraelen, B. Saerens, G. L. Santiago et al.** 2009. Identification and genotyping of bacteria from paired vaginal and rectal samples from pregnant women indicates similarity between vaginal and rectal microflora. BMC.Infect.Dis. **9**:167.

54. **Espy, M. J., J. R. Uhl, L. M. Sloan, S. P. Buckwalter, M. F. Jones, E. A. Vetter et al.** 2006. Real-time PCR in clinical microbiology: applications for routine laboratory testing. Clin.Microbiol.Rev. **19**:165-256.

55. **Evans, J. J., J. F. Bohnsack, P. H. Klesius, A. A. Whiting, J. C. Garcia, C. A. Shoemaker et al.** 2008. Phylogenetic relationships among *Streptococcus agalactiae* isolated from piscine, dolphin, bovine and human sources: a dolphin and piscine lineage associated with a fish epidemic in Kuwait is also associated with human neonatal infections in Japan. J.Med.Microbiol. **57**:1369-1376.

56. **Facklam, R. R., J. F. Padula, E. C. Wortham, R. C. Cooksey, and H. A. Rountree.** 1979. Presumptive identification of group A, B, and D streptococci on agar plate media. J.Clin.Microbiol. **9**:665-672.

57. **Farlow, J., K. L. Smith, J. Wong, M. Abrams, M. Lytle, and P. Keim.** 2001. *Francisella tularensis* Strain Typing Using Multiple-Locus, Variable-Number Tandem Repeat Analysis. Journal of Clinical Microbiology **39**:3186-3192.

58. **Fasola, E., C. Livdahl, and P. Ferrieri.** 1993. Molecular analysis of multiple isolates of the major serotypes of group B streptococci. J.Clin.Microbiol. **31**:2616-2620.

59. **Feil, E. J., B. C. Li, D. M. Aanensen, W. P. Hanage, and B. G. Spratt.** 2004. eBURST: inferring patterns of evolutionary descent among clusters of related bacterial genotypes from multilocus sequence typing data. J.Bacteriol. **186**:1518-1530.

60. **Ferretti, J. J., W. M. McShan, D. Ajdic, D. J. Savic, G. Savic, K. Lyon et al.** 2001. Complete genome sequence of an M1 strain of *Streptococcus pyogenes*. Proc.Natl.Acad.Sci.U.S.A **98**:4658-4663.

61. **Foucault, C., B. La Scola, H. Lindroos, S. G. E. Andersson, and D. Raoult.** 2005. Multispacer Typing Technique for Sequence-Based Typing of *Bartonella quintana*. Journal of Clinical Microbiology **43**:41-48.

62. **Fournier, P. E., Y. O. N. G. ZHU, X. U. E. J. YU, and D. I. D. I. RAOULT.** 2006. Proposal to Create Subspecies of *Rickettsia sibirica* and an Emended Description of *Rickettsia sibirica*. Annals of the New York Academy of Sciences **1078**:597-606.

63. **Fournier, P. E., Y. Zhu, H. Ogata, and D. Raoult.** 2004. Use of Highly Variable Intergenic Spacer Sequences for Multispacer Typing of *Rickettsia conorii* Strains. Journal of Clinical Microbiology **42**:5757-5766.

64. **Foxman, B., B. W. Gillespie, S. D. Manning, and C. F. Marrs.** 2007. Risk factors for group B streptococcal colonization: potential for different transmission systems by capsular type. Ann.Epidemiol. **17**:854-862.

65. **Fraser-Liggett, C. M.** 2005. Insights on biology and evolution from microbial genome sequencing. Genome Research **15**:1603-1610.

66. **G++rtler, V. and V. A. Stanisich.** 1996. New approaches to typing and identification of bacteria using the 16S-23S rDNA spacer region. Microbiology **142**:3-16.

67. **Gascuel, O.** 1997. BIONJ: an improved version of the NJ algorithm based on a simple model of sequence data. Mol.Biol.Evol. **14**:685-695.

68. **Glaser, P., C. Rusniok, C. Buchrieser, F. Chevalier, L. Frangeul, T. Msadek et al.** 2002. Genome sequence of *Streptococcus agalactiae*, a pathogen causing invasive neonatal disease. Mol.Microbiol. **45**:1499-1513.

69. **Glazunova, O., V. Roux, O. Freylikman, Z. Sekeyova, G. Fournous, J. Tyczka et al.** 2005. *Coxiella burnetii* genotyping. Emerg.Infect.Dis. **11**:1211-1217.

70. **Gouy, M., S. Guindon, and O. Gascuel.** 2010. SeaView version 4: A multiplatform graphical user interface for sequence alignment and phylogenetic tree building. Mol.Biol.Evol. **27**:221-224.

71. **Guihot, A., F. Bricaire, and P. Bossi.** 2005. Group B streptococcal meningitis in a patient with horizontal transmission: beware of toothbrushing on Sunday mornings. J.Infect. **50**:240-241.

72. **Guindon, S. and O. Gascuel.** 2003. A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. Syst.Biol. **52**:696-704.

73. **Gulig, P. A., V. de Crecy-Lagard, A. C. Wright, B. Walts, M. Telonis-Scott, and L. M. McIntyre.** 2010. SOLiD sequencing of four *Vibrio vulnificus* genomes enables comparative genomic analysis and identification of candidate clade-specific virulence genes. BMC.Genomics **11**:512.

74. **Gur-Arie, R., C. J. Cohen, Y. Eitan, L. Shelef, E. M. Hallerman, and Y. Kashi.** 2000. Simple sequence repeats in *Escherichia coli:* abundance, distribution, composition, and polymorphism. Genome Res. **10**:62-71.

75. **Gutekunst, H., B. J. Eikmanns, and D. J. Reinscheid.** 2004. The novel fibrinogen-binding protein *FbsB* promotes *Streptococcus agalactiae* invasion into epithelial cells. Infect.Immun. **72**:3495-3504.

76. **Guttormsen, H. K., C. J. Baker, M. H. Nahm, L. C. Paoletti, S. M. Zughaier, M. S. Edwards et al.** 2002. Type III group B streptococcal polysaccharide induces antibodies that cross-react with *Streptococcus pneumoniae* type 14. Infect.Immun. **70**:1724-1738.

77. **Gygax, S. E., J. A. Schuyler, L. E. Kimmel, J. P. Trama, E. Mordechai, and M. E. Adelson.** 2006. Erythromycin and clindamycin resistance in group B streptococcal clinical isolates. Antimicrob.Agents Chemother. **50**:1875-1877.

78. **Haft, D. H., B. J. Loftus, D. L. Richardson, F. Yang, J. A. Eisen, I. T. Paulsen et al.** 2001. TIGRFAMs: a protein family resource for the functional identification of proteins. Nucleic Acids Res. **29**:41-43.

79. **Hall, B. G., G. D. Ehrlich, and F. Z. Hu.** 2010. Pan-genome analysis provides much higher strain typing resolution than multi-locus sequence typing. Microbiology **156**:1060-1068.

80. **Hall, B. G., G. D. Ehrlich, and F. Z. Hu.** 2010. Pan-genome analysis provides much higher strain typing resolution than multi-locus sequence typing. Microbiology **156**:1060-1068.

81. **Hall, N.** 2007. Advanced sequencing technologies and their wider impact in microbiology. Journal of Experimental Biology **210**:1518-1525.

82. **Hannoun, A., M. Shehab, M. T. Khairallah, A. Sabra, R. bi-Rached, T. Bazi et al.** 2009. Correlation between Group B Streptococcal Genotypes, Their Antimicrobial Resistance Profiles, and Virulence Genes among Pregnant Women in Lebanon. Int.J.Microbiol. **2009**:796512.

83. **Harris, S. R., E. J. Feil, M. T. Holden, M. A. Quail, E. K. Nickerson, N. Chantratita et al.** 2010. Evolution of MRSA during hospital transmission and intercontinental spread. Science **327**:469-474.

84. **Harris, S. R., E. J. Feil, M. T. Holden, M. A. Quail, E. K. Nickerson, N. Chantratita et al.** 2010. Evolution of MRSA during hospital transmission and intercontinental spread. Science **327**:469-474.

85. **Health Protection Agency.** 2009. Pyogenic and non pyogenic *streptococcal* bacteraemia, England, Wales and Northern Ireland: 2008. Infection reports **3**.

86. **Heath, P. T., G. Balfour, A. M. Weisner, A. Efstratiou, T. L. Lamagni, H. Tighe et al.** 2004. Group B streptococcal disease in UK and Irish infants younger than 90 days. Lancet **363**:292-294.

87. **Hillenkamp, F. and M. Karas.** 2007. The MALDI Process and Method, p. 1-28. *In* F. Hillenkamp and J. Peter-Katalinic (eds.), MALDI MS: A Practical Guide to Instrumentation, Methods and Applications. Wiley-VCH.

88. **Hiller, N. L., B. Janto, J. S. Hogg, R. Boissy, S. Yu, E. Powell et al.** 2007. Comparative genomic analyses of seventeen *Streptococcus pneumoniae* strains: insights into the pneumococcal supragenome. J.Bacteriol. **189**:8186-8195.

89. **Hoffmaster, A. R., C. C. Fitzgerald, E. Ribot, L. W. Mayer, and T. Popovic.** 2002. Molecular subtyping of *Bacillus anthracis* and the 2001 bioterrorism-associated anthrax outbreak, United States. Emerg.Infect.Dis. **8**:1111-1116.

90. **Holt, K. E., J. Parkhill, C. J. Mazzoni, P. Roumagnac, F. X. Weill, I. Goodhead et al.** 2008. High-throughput sequencing provides insights into genome variation and evolution in *Salmonella Typhi*. Nat.Genet. **40**:987-993.

91. **Honisch, C., Y. Chen, C. Mortimer, C. Arnold, O. Schmidt, B. D. van den et al.** 2007. Automated comparative sequence analysis by base-specific cleavage and mass spectrometry for nucleic acid-based microbial typing. Proc.Natl.Acad.Sci.U.S.A **104**:10649-10654.

92. **Hood, D. W., M. E. Deadman, M. P. Jennings, M. Bisercic, R. D. Fleischmann, J. C. Venter et al.** 1996. DNA repeats identify novel virulence genes in *Haemophilus influenzae*. Proc.Natl.Acad.Sci.U.S.A **93**:11121-11125.

93. **Hynes, W. L., L. Hancock, and J. J. Ferretti.** 1995. Analysis of a second bacteriophage hyaluronidase gene from *Streptococcus pyogenes*: evidence for a third hyaluronidase involved in extracellular enzymatic activity. Infect.Immun. **63**:3015-3020.

94. **Ibrahim, Y. M., A. R. Kerr, J. McCluskey, and T. J. Mitchell.** 2004. Control of virulence by the two-component system *CiaR/H* is mediated via *HtrA*, a major virulence factor of *Streptococcus pneumoniae*. J.Bacteriol. **186**:5258-5266.

95. **Ippolito, D. L., W. A. James, D. Tinnemore, R. R. Huang, M. J. Dehart, J. Williams et al.** 2010. Group B streptococcus serotype prevalence in reproductive-age women at a tertiary care military medical center relative to global serotype distribution. BMC.Infect.Dis. **10**:336.

96. **Jacobsson, K.** 2003. A novel family of fibrinogen-binding proteins in *Streptococcus agalactiae*. Vet.Microbiol. **96**:103-113.

97. **Janulczyk, R., V. Masignani, D. Maione, H. Tettelin, G. Grandi, and J. L. Telford.** 2010. Simple sequence repeats and genome plasticity in *Streptococcus agalactiae*. J.Bacteriol. **192**:3990-4000.

98. **Jerlstrom, P. G., S. R. Talay, P. Valentin-Weigand, K. N. Timmis, and G. S. Chhatwal.** 1996. Identification of an immunoglobulin A binding motif located in the beta-antigen of the c protein complex of group B streptococci. Infect.Immun. **64**:2787-2793.

99. **Johansson, A., J. Farlow, P. +. Larsson, M. Dukerich, E. Chambers, M. Bystr+Âm et al.** 2004. Worldwide Genetic Relationships among *Francisella tularensis* Isolates Determined by Multiple-Locus Variable-Number Tandem Repeat Analysis. Journal of Bacteriology **186**:5808-5818.

100. **Johri, A. K., L. C. Paoletti, P. Glaser, M. Dua, P. K. Sharma, G. Grandi et al.** 2006. Group B Streptococcus: global incidence and vaccine development. Nat.Rev.Microbiol. **4**:932-942.

101. **Jolley, K. A., M. S. Chan, and M. C. Maiden.** 2004. mlstdbNet - distributed multi-locus sequence typing (MLST) databases. BMC.Bioinformatics. **5**:86.

102. **Jolley, K. A., E. J. Feil, M. S. Chan, and M. C. Maiden.** 2001. Sequence type analysis and recombinational tests (START). Bioinformatics. **17**:1230-1231.

103. **Jones, N., J. F. Bohnsack, S. Takahashi, K. A. Oliver, M. S. Chan, F. Kunst et al.** 2003. Multilocus sequence typing system for group B streptococcus. J.Clin.Microbiol. **41**:2530-2536.

104. **Jones, N., K. A. Oliver, J. Barry, R. M. Harding, N. Bisharat, B. G. Spratt et al.** 2006. Enhanced invasiveness of bovine-derived neonatal sequence type 17 group B streptococcus is independent of capsular serotype. Clin.Infect.Dis. **42**:915-924.

105. **Jordan, J. A., A. R. Butchko, and M. B. Durso**. 2005. Use of pyrosequencing of 16S rRNA fragments to differentiate between bacteria responsible for neonatal sepsis. J.Mol.Diagn. **7**:105-110.

106. **Jordan, J. A. and M. B. Durso**. 2005. Real-time polymerase chain reaction for detecting bacterial DNA directly from blood of neonates being evaluated for sepsis. J.Mol.Diagn. **7**:575-581.

107. **Jordan, J. A., M. B. Durso, A. R. Butchko, J. G. Jones, and B. S. Brozanski**. 2006. Evaluating the near-term infant for early onset sepsis: progress and challenges to consider with 16S rDNA polymerase chain reaction testing. J.Mol.Diagn. **8**:357-363.

108. **Jordan, P., L. A. Snyder, and N. J. Saunders.** 2003. Diversity in coding tandem repeats in related *Neisseria* spp. BMC.Microbiol. **3**:23.

109. **Jurinke, C., P. Oeth, and D. van den Boom**. 2004. MALDI-TOF mass spectrometry: a versatile tool for high-performance DNA analysis. Mol.Biotechnol. **26**:147-164.

110. **Ke, D., C. Menard, F. J. Picard, M. Boissinot, M. Ouellette, P. H. Roy et al.** 2000. Development of conventional and real-time PCR assays for the rapid detection of group B streptococci. Clin.Chem. **46**:324-331.

111. **Keim, P., L. B. Price, A. M. Klevytska, K. L. Smith, J. M. Schupp, R. Okinaka et al.** 2000. Multiple-Locus Variable-Number Tandem Repeat Analysis Reveals Genetic Relationships within *Bacillus anthracis*. Journal of Bacteriology **182**:2928-2936.

112. **Kilian, M.** 2005. *Streptococcus* and *Lactobacillus*, p. 833-867. *In* S. P. F. G. Borriello and P. R. Murray (eds.), Topley & Wilsons Microbiology & Microbial Infections: Bacteriology.

113. **Klevytska, A. M., L. B. Price, J. M. Schupp, P. L. Worsham, J. Wong, and P. Keim**. 2001. Identification and Characterization of Variable-Number Tandem Repeats in the *Yersinia pestis* Genome. Journal of Clinical Microbiology **39**:3179-3185.

114. **Kong, F., S. Gowan, D. Martin, G. James, and G. L. Gilbert**. 2002. Molecular profiles of group B streptococcal surface protein antigen genes: relationship to molecular serotypes. J.Clin.Microbiol. **40**:620-626.

115. **Kong, F., S. Gowan, D. Martin, G. James, and G. L. Gilbert**. 2002. Serotype identification of group B streptococci by PCR and sequencing. J.Clin.Microbiol. **40**:216-226.

116. **Kong, F., L. M. Lambertsen, H. C. Slotved, D. Ko, H. Wang, and G. L. Gilbert**. 2008. Use of phenotypic and molecular serotype identification methods to characterise previously non-serotypeable group B streptococci (GBS). J.Clin.Microbiol.

117. **Kong, F., D. Martin, G. James, and G. L. Gilbert**. 2003. Towards a genotyping system for *Streptococcus agalactiae* (group B streptococcus): use of mobile genetic elements in Australasian invasive isolates. J.Med.Microbiol. **52**:337-344.

118. **Konstantinidis, K. T., A. Ramette, and J. M. Tiedje.** 2006. The bacterial species definition in the genomic era. Philos.Trans.R.Soc.Lond B Biol.Sci. **361**:1929-1940.

119. **Konstantinidis, K. T., A. Ramette, and J. M. Tiedje.** 2006. Toward a more robust assessment of intraspecies diversity, using fewer genetic markers. Appl.Environ.Microbiol. **72**:7286-7293.

120. **Konstantinidis, K. T. and J. M. Tiedje.** 2005. Genomic insights that advance the species definition for prokaryotes. Proc.Natl.Acad.Sci.U.S.A **102**:2567-2572.

121. **Konstantinidis, K. T. and J. M. Tiedje.** 2005. Towards a genome-based taxonomy for prokaryotes. J.Bacteriol. **187**:6258-6264.

122. **Lachenauer, C. S., R. Creti, J. L. Michel, and L. C. Madoff.** 2000. Mosaicism in the alpha-like protein genes of group B streptococci. Proc.Natl.Acad.Sci.U.S.A **97**:9630-9635.

123. **Lachenauer, C. S., D. L. Kasper, J. Shimada, Y. Ichiman, H. Ohtsuka, M. Kaku et al.** 1999. Serotypes VI and VIII predominate among group B streptococci isolated from pregnant Japanese women. J.Infect.Dis. **179**:1030-1033.

124. **Laing, C., C. Buchanan, E. N. Taboada, Y. Zhang, A. Kropinski, A. Villegas et al.** 2010. Pan-genome sequence analysis using Panseq: an online tool for the rapid analysis of core and accessory genomic regions. BMC.Bioinformatics. **11**:461.

125. **Lamy, M. C., S. Dramsi, A. Billoet, H. Reglier-Poupet, A. Tazi, J. Raymond et al.** 2006. Rapid detection of the "highly virulent" group B Streptococcus ST-17 clone. Microbes.Infect. **8**:1714-1722.

126. **Lancefield, R. C.** 1934. A serological differentiation of specific types of bovine hemolytic streptococci (Group B). Journal of Experimental Medicine **59**:441-458.

127. **Lauer, P., C. D. Rinaudo, M. Soriani, I. Margarit, D. Maione, R. Rosini et al.** 2005. Genome analysis reveals pili in Group B Streptococcus. Science **309**:105.

128. **Law, M. R., G. Palomaki, Z. Alfirevic, R. Gilbert, P. Heath, C. McCartney et al.** 2005. The prevention of neonatal group B streptococcal disease: a report by a working group of the Medical Screening Society. J.Med.Screen. **12**:60-68.

129. **Le Fleche, P., M. Fabre, F. Denoeud, J. L. Koeck, and G. Vergnaud.** 2002. High resolution, on-line identification of strains from the *Mycobacterium tuberculosis* complex based on tandem repeat typing. BMC Microbiology **2**:37.

130. **Lee, N. Y., J. J. Yan, J. J. Wu, H. C. Lee, K. H. Liu, and W. C. Ko.** 2005. Group B streptococcal soft tissue infections in non-pregnant adults. Clin.Microbiol.Infect. **11**:577-579.

131. **Lee, S., K. H. Roh, C. K. Kim, D. Yong, J. Y. Choi, J. W. Lee et al.** 2008. A Case of Necrotizing Fasciitis Due to *Streptococcus agalactiae*, *Arcanobacterium haemolyticum*, and *Finegoldia magna* in a Dog-bitten Patient with Diabetes. Korean J.Lab Med. **28**:191-195.

132. **Lefebure, T. and M. J. Stanhope.** 2007. Evolution of the core and pan-genome of *Streptococcus*: positive selection, recombination, and genome composition. Genome Biol. **8**:R71.

133. **Levinson, G. and G. A. Gutman**. 1987. Slipped-strand mispairing: a major mechanism for DNA sequence evolution. Mol.Biol.Evol. **4**:203-221.

134. **Li, J. J., G. L. Pei, H. X. Pang, A. Bilderbeck, S. S. Chen, and S. H. Tao**. 2006. A new method for RAPD primers selection based on primer bias in nucleotide sequence data. Journal of Biotechnology **126**:415-423.

135. **Li, W. and A. Godzik**. 2006. Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. Bioinformatics. **22**:1658-1659.

136. **Li, W., B. B. Chomel, S. Maruyama, L. Guptil, A. Sander, D. Raoult et al.** 2006. Multispacer Typing To Study the Genotypic Distribution of *Bartonella henselae* Populations. Journal of Clinical Microbiology **44**:2499-2506.

137. **Li, W., F. Fenollar, J. M. Rolain, P. E. Fournier, G. E. Feurle, C. Muller et al.** 2008. Genotyping reveals a wide heterogeneity of *Tropheryma whipplei*. Microbiology **154**:521-527.

138. **Li, W., D. Raoult, and P. E. Fournier**. 2009. Bacterial strain typing in the genomic era. FEMS Microbiology Reviews **33**:892-916.

139. **Lindstedt, B. A.** 2005. Multiple-locus variable number tandem repeats analysis for genetic fingerprinting of pathogenic bacteria. Electrophoresis **26**:2567-2582.

140. **Lista, F., G. Faggioni, S. Valjevac, A. Ciammaruconi, J. Vaissaire, C. le Doujet et al.** 2006. Genotyping of *Bacillus anthracis* strains based on automated capillary 25-loci Multiple Locus Variable-Number Tandem Repeats Analysis. BMC Microbiology **6**:33.

141. **Luan, S. L., M. Granlund, M. Sellin, T. Lagergard, B. G. Spratt, and M. Norgren**. 2005. Multilocus sequence typing of Swedish invasive group B streptococcus isolates indicates a neonatally associated genetic lineage and capsule switching. J.Clin.Microbiol. **43**:3727-3733.

142. **Lukjancenko, O., T. M. Wassenaar, and D. W. Ussery**. 2010. Comparison of 61 sequenced *Escherichia coli* genomes. Microb.Ecol. **60**:708-720.

143. **Luna, R. A., L. R. Fasciano, S. C. Jones, B. L. Boyanton, Jr., T. T. Ton, and J. Versalovic**. 2007. DNA pyrosequencing-based bacterial pathogen identification in a pediatric hospital setting. J.Clin.Microbiol. **45**:2985-2992.

144. **Lupski, J. R. and G. M. Weinstock**. 1992. Short, interspersed repetitive DNA sequences in prokaryotic genomes. J.Bacteriol. **174**:4525-4529.

145. **Ma, L., S. Taylor, J. S. Jensen, L. Myers, R. Lillis, and D. H. Martin**. 2008. Short tandem repeat sequences in the *Mycoplasma genitalium* genome and their use in a multilocus genotyping system. BMC.Microbiol. **8**:130.

146. **Maiden, M. C., J. A. Bygraves, E. Feil, G. Morelli, J. E. Russell, R. Urwin et al.** 1998. Multilocus sequence typing: a portable approach to the identification of clones within populations of pathogenic microorganisms. Proc.Natl.Acad.Sci.U.S.A **95**:3140-3145.

147. **Maione, D., I. Margarit, C. D. Rinaudo, V. Masignani, M. Mora, M. Scarselli et al.** 2005. Identification of a universal Group B streptococcus vaccine by multiple genome screen. Science **309**:148-150.

148. **Maisey, H. C., M. Hensler, V. Nizet, and K. S. Doran**. 2007. Group B streptococcal pilus proteins contribute to adherence to and invasion of brain microvascular endothelial cells. J.Bacteriol. **189**:1464-1467.

149. **Manning, S. D., M. A. Lewis, A. C. Springman, E. Lehotzky, T. S. Whittam, and H. D. Davies**. 2008. Genotypic Diversity and Serotype Distribution of Group B Streptococcus Isolated from Women Before and After Delivery. Clin.Infect.Dis.

150. **Manning, S. D., A. C. Springman, E. Lehotzky, M. A. Lewis, T. S. Whittam, and H. D. Davies**. 2009. Multilocus sequence types associated with neonatal group B streptococcal sepsis and meningitis in Canada. J.Clin.Microbiol. **47**:1143-1148.

151. **Manning, S. D., A. C. Springman, A. D. Million, N. R. Milton, S. E. McNamara, P. A. Somsel et al.** 2010. Association of Group B Streptococcus colonization and bovine exposure: a prospective cross-sectional cohort study. PLoS.ONE. **5**:e8795.

152. **Marchaim, D., S. Efrati, R. Melamed, L. Gortzak-Uzan, K. Riesenberg, R. Zaidenstein et al.** 2006. Clonal variability of group B Streptococcus among different groups of carriers in southern Israel. Eur.J.Clin.Microbiol.Infect.Dis. **25**:443-448.

153. **Martin, P., d. van, V, N. Mouchel, A. C. Jeffries, D. W. Hood, and E. R. Moxon**. 2003. Experimentally revised repertoire of putative contingency loci in *Neisseria meningitidis* strain MC58: evidence for a novel mechanism of phase variation. Mol.Microbiol. **50**:245-257.

154. **Martinez, G., J. Harel, R. Higgins, S. Lacouture, D. Daignault, and M. Gottschalk**. 2000. Characterization of *Streptococcus agalactiae* isolates of bovine and human origin by randomly amplified polymorphic DNA analysis. J.Clin.Microbiol. **38**:71-78.

155. **Martins, E. R., J. Melo-Cristino, and M. Ramirez**. 2010. Evidence for rare capsular switching in *Streptococcus agalactiae*. J.Bacteriol. **192**:1361-1369.

156. **Martins, E. R., M. A. Pessanha, M. Ramirez, and J. Melo-Cristino**. 2007. Analysis of group B streptococcal isolates from infants and pregnant women in Portugal revealing two lineages with enhanced invasiveness. J.Clin.Microbiol. **45**:3224-3229.

157. **Mashayekhi, F. and M. Ronaghi**. 2007. Analysis of read length limiting factors in Pyrosequencing chemistry. Anal.Biochem. **363**:275-287.

158. **McClelland, M., R. Jones, Y. Patel, and M. Nelson**. 1987. Restriction endonucleases for pulsed field mapping of bacterial genomes. Nucleic Acids Res. **15**:5985-6005.

159. **Medini, D., C. Donati, H. Tettelin, V. Masignani, and R. Rappuoli**. 2005. The microbial pan-genome. Curr.Opin.Genet.Dev. **15**:589-594.

160. **Mellmann, A., D. Harmsen, C. A. Cummings, E. B. Zentz, S. R. Leopold, A. Rico et al.** 2011. Prospective genomic characterization of the German enterohemorrhagic *Escherichia coli* O104:H4 outbreak by rapid next generation sequencing technology. PLoS.ONE. **6**:e22751.

161. **Mereghetti, L., I. Sitkiewicz, N. M. Green, and J. M. Musser**. 2008. Extensive adaptive changes occur in the transcriptome of *Streptococcus agalactiae* (group B streptococcus) in response to incubation with human blood. PLoS.ONE. **3**:e3143.

162. **Monot, M., N. Honor+®, C. Bali+¿re, B. Ji, S. Sow, P. J. Brennan et al.** 2008. Are Variable-Number Tandem Repeats Appropriate for Genotyping *Mycobacterium leprae*? Journal of Clinical Microbiology **46**:2291-2297.

163. **Morelli, G., Y. Song, C. J. Mazzoni, M. Eppinger, P. Roumagnac, D. M. Wagner et al.** 2010. *Yersinia pestis* genome sequencing identifies patterns of global phylogenetic diversity. Nat.Genet.

164. **Mount, D.** 2004. Bioinformatics: Sequence and Genome Analysis. Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY.

165. **Moxon, R., C. Bayliss, and D. Hood**. 2006. Bacterial contingency loci: the role of simple sequence DNA repeats in bacterial adaptation. Annu.Rev.Genet. **40**:307-333.

166. **Natarajan, G., Y. R. Johnson, F. Zhang, K. M. Chen, and M. J. Worsham**. 2006. Real-time polymerase chain reaction for the rapid detection of group B streptococcal colonization in neonates. Pediatrics **118**:14-22.

167. **Nizet, V., P. Ferrieri, and C. E. Rubens**. 2000. Molecular pathogenesis of group B streptococcal disease in newborns., p. 180-221. *In* D. L. Stevens and E. L. Kaplan (eds.), Clinical Aspects, Microbiology and Molecular Pathogenesis. Oxford University Press, New York.

168. **Nizet, V., R. L. Gibson, E. Y. Chi, P. E. Framson, M. Hulse, and C. E. Rubens**. 1996. Group B streptococcal beta-hemolysin expression is associated with injury of lung epithelial cells. Infect.Immun. **64**:3818-3826.

169. **Nygren, M., E. Reizenstein, M. Ronaghi, and J. Lundeberg**. 2000. Polymorphism in the pertussis toxin promoter region affecting the DNA-based diagnosis of *Bordetella* infection. J.Clin.Microbiol. **38**:55-60.

170. **Ogata, H., S. p. Audic, V. r. Barbe, F. Artiguenave, P. E. Fournier, D. Raoult et al.** 2000. Selfish DNA in Protein-Coding Genes of Rickettsia. Science **290**:347-350.

171. **Orsi, R. H., B. M. Bowen, and M. Wiedmann**. 2010. Homopolymeric tracts represent a general regulatory mechanism in prokaryotes. BMC.Genomics **11**:102.

172. **Pallen, M. J., N. J. Loman, and C. W. Penn**. 2010. High-throughput sequencing and clinical microbiology: progress, opportunities and challenges. Curr.Opin.Microbiol. **13**:625-631.

173. **Pancholi, V. and V. A. Fischetti**. 1998. alpha-enolase, a novel strong plasmin(ogen) binding protein on the surface of pathogenic streptococci. J.Biol.Chem. **273**:14503-14515.

174. **Paoletti, L. C., H. K. Guttormsen, M. S. Christian, A. M. Hoberman, and P. McInnes**. 2008. Neither antibody to a group B streptococcal conjugate vaccine nor the vaccine itself is teratogenic in rabbits. Hum.Vaccin. **4**:435-443.

175. **Paoletti, L. C. and L. C. Madoff**. 2002. Vaccines to prevent neonatal GBS infection. Semin.Neonatol. **7**:315-323.

176. **Paoletti, L. C., M. R. Wessels, A. K. Rodewald, A. A. Shroff, H. J. Jennings, and D. L. Kasper**. 1994. Neonatal mouse protection against infection with multiple group B streptococcal (GBS) serotypes by maternal immunization with a tetravalent GBS polysaccharide-tetanus toxoid conjugate vaccine. Infect.Immun. **62**:3236-3243.

177. **Paredes, A., P. Wong, E. O. Mason, Jr., L. H. Taber, and F. F. Barrett.** 1977. Nosocomial transmission of group B Streptococci in a newborn nursery. Pediatrics **59**:679-682.

178. **Parkhill, J., B. W. Wren, K. Mungall, J. M. Ketley, C. Churcher, D. Basham et al.** 2000. The genome sequence of the food-borne pathogen *Campylobacter jejuni* reveals hypervariable sequences. Nature **403**:665-668.

179. **Pearson, T., P. Giffard, S. Beckstrom-Sternberg, R. Auerbach, H. Hornstra, A. Tuanyok et al.** 2009. Phylogeographic reconstruction of a bacterial species with high levels of lateral gene transfer. BMC Biol. **7**:78.

180. **Picard, F. J. and M. G. Bergeron.** 2004. Laboratory detection of group B Streptococcus for prevention of perinatal disease. Eur.J.Clin.Microbiol.Infect.Dis. **23**:665-671.

181. **Posada, D.** 2008. jModelTest: phylogenetic model averaging. Mol.Biol.Evol. **25**:1253-1256.

182. **Posada, D. and K. A. Crandall.** 1998. MODELTEST: testing the model of DNA substitution. Bioinformatics. **14**:817-818.

183. **Pritzlaff, C. A., J. C. Chang, S. P. Kuo, G. S. Tamura, C. E. Rubens, and V. Nizet.** 2001. Genetic basis for the beta-haemolytic/cytolytic activity of group B Streptococcus. Mol.Microbiol. **39**:236-247.

184. **Puopolo, K. M. and L. C. Madoff.** 2003. Upstream short sequence repeats regulate expression of the alpha C protein of group B Streptococcus. Mol.Microbiol. **50**:977-991.

185. **Quail, M. A., M. Smith, P. Coupland, T. D. Otto, S. R. Harris, T. R. Connor et al.** 2012. A tale of three next generation sequencing platforms: comparison of Ion Torrent, Pacific Biosciences and Illumina MiSeq sequencers. BMC Genomics **13**:341.

186. **Radtke, A., B. A. Lindstedt, J. E. Afset, and K. Bergh.** 2010. Rapid multiple-locus variant-repeat assay (MLVA) for genotyping of *Streptococcus agalactiae*. J.Clin.Microbiol. **48**:2502-2508.

187. **Rajagopal, L.** 2009. Understanding the regulation of Group B Streptococcal virulence factors. Future.Microbiol. **4**:201-221.

188. **Rajagopal, L., A. Vo, A. Silvestroni, and C. E. Rubens.** 2005. Regulation of purine biosynthesis by a eukaryotic-type kinase in *Streptococcus agalactiae*. Mol.Microbiol. **56**:1329-1346.

189. **Rajagopal, L., A. Vo, A. Silvestroni, and C. E. Rubens.** 2006. Regulation of cytotoxin expression by converging eukaryotic-type and two-component signalling mechanisms in *Streptococcus agalactiae*. Mol.Microbiol. **62**:941-957.

190. **Rallu, F., P. Barriga, C. Scrivo, V. Martel-Laferriere, and C. Laferriere.** 2006. Sensitivities of antigen detection and PCR assays greatly increased compared to that of the standard culture method for screening for group B streptococcus carriage in pregnant women. J.Clin.Microbiol. **44**:725-728.

191. **Ramaswamy, S. V., P. Ferrieri, A. E. Flores, and L. C. Paoletti.** 2006. Molecular characterization of nontypeable group B streptococcus. J.Clin.Microbiol. **44**:2398-2403.

192. **Rinaudo, C. D., J. L. Telford, R. Rappuoli, and K. L. Seib.** 2009. Vaccinology in the genome era. J.Clin.Invest **119**:2515-2525.

193. **Rioux, S., D. Martin, H. W. Ackermann, J. Dumont, J. Hamel, and B. R. Brodeur**. 2001. Localization of surface immunogenic protein on group B streptococcus. Infect.Immun. **69**:5162-5165.

194. **Rolland, K., C. Marois, V. Siquier, B. Cattier, and R. Quentin**. 1999. Genetic features of *Streptococcus agalactiae* strains causing severe neonatal infections, as revealed by pulsed-field gel electrophoresis and *hylB* gene analysis. J.Clin.Microbiol. **37**:1892-1898.

195. **Rosini, R., C. D. Rinaudo, M. Soriani, P. Lauer, M. Mora, D. Maione et al.** 2006. Identification of novel genomic islands coding for antigenic pilus-like structures in *Streptococcus agalactiae*. Mol.Microbiol. **61**:126-141.

196. **Royal College of Obstetricians and Gynaecologists**. 2003. Prevention of Early Onset Neonatal Group B Streptococcal Disease, p. RCOG Guideline No. 36.

197. **Rozhdestvenskaya, A. S., A. A. Totolian, and A. V. Dmitriev**. 2010. Inactivation of DNA-binding response regulator Sak189 abrogates beta-antigen expression and affects virulence of *Streptococcus agalactiae*. PLoS.ONE. **5**:e10212.

198. **Rusk, N.** 2009. Focus on next-generation sequencing data analysis. Forward. Nat.Methods **6**:S1.

199. **Sadeghifard, N., V. G++rtler, M. Beer, and R. J. Seviour**. 2006. The Mosaic Nature of Intergenic 16S-23S rRNA Spacer Regions Suggests rRNA Operon Copy Number Variation in *Clostridium difficile* Strains. Applied and Environmental Microbiology **72**:7311-7323.

200. **Saitou, N. and M. Nei**. 1987. The neighbor-joining method: a new method for reconstructing phylogenetic trees. Mol.Biol.Evol. **4**:406-425.

201. **Salzberg, S. L., A. L. Delcher, S. Kasif, and O. White**. 1998. Microbial gene identification using interpolated Markov models. Nucleic Acids Res. **26**:544-548.

202. **Samen, U. M., B. J. Eikmanns, and D. J. Reinscheid**. 2006. The transcriptional regulator *RovS* controls the attachment of *Streptococcus agalactiae* to human epithelial cells and the expression of virulence genes. Infect.Immun. **74**:5625-5635.

203. **Sanger, F. and A. R. Coulson**. 1975. A rapid method for determining sequences in DNA by primed synthesis with DNA polymerase. J.Mol.Biol. **94**:441-448.

204. **Sanger, F., S. Nicklen, and A. R. Coulson**. 1977. DNA sequencing with chain-terminating inhibitors. Proc.Natl.Acad.Sci.U.S.A **74**:5463-5467.

205. **Saunders, N. J., J. F. Peden, D. W. Hood, and E. R. Moxon**. 1998. Simple sequence repeats in the *Helicobacter pylori* genome. Mol.Microbiol. **27**:1091-1098.

206. **Schouls, L. M., E. C. Spalburg, L. M. van, X. W. Huijsdens, G. N. Pluister, M. G. van Santen-Verheuvel et al.** 2009. Multiple-locus variable number tandem repeat analysis of *Staphylococcus aureus*: comparison with pulsed-field gel electrophoresis and *spa*-typing. PLoS.ONE. **4**:e5082.

207. **Schrag, S., R. Gorwitz, K. Fultz-Butts, and A. Schuchat**. 2002. Prevention of perinatal group B streptococcal disease. Revised guidelines from CDC. MMWR Recomm.Rep. **51**:1-22.

208. **Schubert, A., K. Zakikhany, M. Schreiner, R. Frank, B. Spellerberg, B. J. Eikmanns et al.** 2002. A fibrinogen receptor from group B Streptococcus interacts with fibrinogen by repetitive units with novel ligand binding sites. Mol.Microbiol. **46**:557-569.

209. **Schuchat, A.** 2001. Group B streptococcal disease: from trials and tribulations to triumph and trepidation. Clin.Infect.Dis. **33**:751-756.

210. **Schwartz, D. C. and C. R. Cantor**. 1984. Separation of yeast chromosome-sized DNAs by pulsed field gradient gel electrophoresis. Cell **37**:67-75.

211. **Seepersaud, R., S. B. Hanniffy, P. Mayne, P. Sizer, P. R. Le, and J. M. Wells**. 2005. Characterization of a novel leucine-rich repeat protein antigen from group B streptococci that elicits protective immunity. Infect.Immun. **73**:1671-1683.

212. **Sela, S., A. Aviv, A. Tovi, I. Burstein, M. G. Caparon, and E. Hanski**. 1993. Protein F: an adhesin of *Streptococcus pyogenes* binds fibronectin via two distinct domains. Mol.Microbiol. **10**:1049-1055.

213. **Sitkiewicz, I., N. M. Green, N. Guo, A. M. Bongiovanni, S. S. Witkin, and J. M. Musser.** 2009. Transcriptome adaptation of group B Streptococcus to growth in human amniotic fluid. PLoS.ONE. **4**:e6114.

214. **Sitkiewicz, I. and J. M. Musser**. 2009. Analysis of growth-phase regulated genes in *Streptococcus agalactiae* by global transcript profiling. BMC.Microbiol. **9**:32.

215. **Slotved, H. C., J. Elliott, T. Thompson, and H. B. Konradsen**. 2003. Latex assay for serotyping of group B Streptococcus isolates. J.Clin.Microbiol. **41**:4445-4447.

216. **Slotved, H. C., F. Kong, L. Lambertsen, S. Sauer, and G. L. Gilbert**. 2007. Serotype IX, a Proposed New *Streptococcus agalactiae* Serotype. J.Clin.Microbiol. **45**:2929-2936.

217. **Snyder, L. A., S. A. Butcher, and N. J. Saunders**. 2001. Comparative whole-genome analyses reveal over 100 putative phase-variable genes in the pathogenic *Neisseria* spp. Microbiology **147**:2321-2332.

218. **Sober, E.** 1983. Parsimony in Systematics: Philosophical Issues. Annual Review of Ecology and Systematics **14**:335-357.

219. **Sorensen, U. B., K. Poulsen, C. Ghezzo, I. Margarit, and M. Kilian**. 2010. Emergence and Global Dissemination of Host-Specific *Streptococcus agalactiae* Clones. MBio. **1**.

220. **Spellerberg, B., E. Rozdzinski, S. Martin, J. Weber-Heynemann, N. Schnitzler, R. Lutticken et al.** 1999. *Lmb*, a protein with similarities to the *Lral* adhesin family, mediates attachment of *Streptococcus agalactiae* to human laminin. Infect.Immun. **67**:871-878.

221. **Springman, A. C., D. W. Lacher, G. Wu, N. Milton, T. S. Whittam, H. D. Davies et al.** 2009. Selection, recombination and virulence gene diversity among group B streptococcal genotypes. J.Bacteriol.

222. **Sreenu, V. B., P. Kumar, J. Nagaraju, and H. A. Nagarajaram**. 2006. Microsatellite polymorphism across the *M. tuberculosis* and *M. bovis* genomes: implications on genome evolution and plasticity. BMC.Genomics **7**:78.

223. **Stackebrandt, E. and B. M. GOEBEL**. 1994. Taxonomic Note: A Place for DNA-DNA Reassociation and 16S rRNA Sequence Analysis in the Present Species Definition in Bacteriology. International Journal of Systematic Bacteriology **44**:846-849.

224. **Stalhammar-Carlemalm, M., L. Stenberg, and G. Lindahl.** 1993. Protein rib: a novel group B streptococcal cell surface protein that confers protective immunity and is expressed by most strains causing invasive infections. J.Exp.Med. **177**:1593-1603.

225. **Stanssens, P., M. Zabeau, G. Meersseman, G. Remes, Y. Gansemans, N. Storm et al.** 2004. High-throughput MALDI-TOF discovery of genomic sequence polymorphisms. Genome Res. **14**:126-133.

226. **Stoll, B. J., N. Hansen, A. A. Fanaroff, L. L. Wright, W. A. Carlo, R. A. Ehrenkranz et al.** 2002. Late-onset sepsis in very low birth weight neonates: the experience of the NICHD Neonatal Research Network. Pediatrics **110**:285-291.

227. **Straka, M., C. W. Dela, C. Blackmon, O. Johnson, S. Stassen, D. Streitman et al.** 2004. Rapid detection of group B streptococcus and *Escherichia coli* in amniotic fluid using real-time fluorescent PCR. Infect.Dis.Obstet.Gynecol. **12**:109-114.

228. **Sun, Y., F. Kong, Z. Zhao, and G. L. Gilbert.** 2005. Comparison of a 3-set genotyping system with multilocus sequence typing for *Streptococcus agalactiae* (Group B Streptococcus). J.Clin.Microbiol. **43**:4704-4707.

229. **Telford, J. L., M. A. Barocchi, I. Margarit, R. Rappuoli, and G. Grandi.** 2006. Pili in gram-positive pathogens. Nat.Rev.Microbiol. **4**:509-519.

230. **Tenover, F. C., R. D. Arbeit, R. V. Goering, P. A. Mickelsen, B. E. Murray, D. H. Persing et al.** 1995. Interpreting chromosomal DNA restriction patterns produced by pulsed-field gel electrophoresis: criteria for bacterial strain typing. J.Clin.Microbiol. **33**:2233-2239.

231. **Tettelin, H., V. Masignani, M. J. Cieslewicz, C. Donati, D. Medini, N. L. Ward et al.** 2005. Genome analysis of multiple pathogenic isolates of *Streptococcus agalactiae*: implications for the microbial "pan-genome". Proc.Natl.Acad.Sci.U.S.A **102**:13950-13955.

232. **Tettelin, H., V. Masignani, M. J. Cieslewicz, J. A. Eisen, S. Peterson, M. R. Wessels et al.** 2002. Complete genome sequence and comparative genomic analysis of an emerging human pathogen, serotype V *Streptococcus agalactiae*. Proc.Natl.Acad.Sci.U.S.A **99**:12391-12396.

233. **Tettelin, H., K. E. Nelson, I. T. Paulsen, J. A. Eisen, T. D. Read, S. Peterson et al.** 2001. Complete genome sequence of a virulent isolate of *Streptococcus pneumoniae*. Science **293**:498-506.

234. **Touzain, F., E. Denamur, C. Medigue, V. Barbe, K. M. El, and M. A. Petit.** 2010. Small variable segments constitute a major type of diversity of bacterial genomes at the species level. Genome Biol. **11**:R45.

235. **Tumapa, S., M. T. Holden, M. Vesaratchavest, V. Wuthiekanun, D. Limmathurotsakul, W. Chierakul et al.** 2008. *Burkholderia pseudomallei* genome plasticity associated with genomic island variation. BMC.Genomics **9**:190.

236. **U'Ren, J., J. Schupp, T. Pearson, H. Hornstra, C. Friedman, K. Smith et al.** 2007. Tandem repeat regions within the *Burkholderia pseudomallei* genome and their application for high resolution genotyping. BMC Microbiology **7**:23.

237. **Uchiyama, I.** 2008. Multiple genome alignment for identifying the core structure among moderately related microbial genomes. BMC.Genomics **9**:515.

238. **Uhl, J. R., E. A. Vetter, K. L. Boldt, B. W. Johnston, K. D. Ramin, M. J. Adams et al.** 2005. Use of the Roche LightCycler Strep B assay for detection of group B Streptococcus from vaginal and rectal swabs. J.Clin.Microbiol. **43**:4046-4051.

239. **Ulett, K. B., W. H. Benjamin, Jr., F. Zhuo, M. Xiao, F. Kong, G. L. Gilbert et al.** 2009. Diversity of group B streptococcus serotypes causing urinary tract infection in adults. J.Clin.Microbiol. **47**:2055-2060.

240. **van Belkum, A., S. Scherer, L. van Alphen, and H. Verbrugh.** 1998. Short-Sequence DNA Repeats in Prokaryotic Genomes. Microbiology and Molecular Biology Reviews **62**:275-293.

241. **van den Berg, R. J., I. Schaap, K. E. Templeton, C. H. W. Klaassen, and E. J. Kuijper.** 2007. Typing and Subtyping of *Clostridium difficile* Isolates by Using Multiple-Locus Variable-Number Tandem-Repeat Analysis. Journal of Clinical Microbiology **45**:1024-1028.

242. **van der Woude, M. W. and A. J. Baumler.** 2004. Phase and antigenic variation in bacteria. Clin.Microbiol.Rev. **17**:581-611, table.

243. **van, E. E., R. Yahiaoui, C. Visser, P. Oostvogel, A. Muller, Y. R. Ho et al.** 2009. Epidemiology of and prenatal molecular distinction between invasive and colonizing group B streptococci in The Netherlands and Taiwan. Eur.J.Clin.Microbiol.Infect.Dis. **28**:921-928.

244. **van-der Mee-Marquet, N., A. S. Domelier, L. Mereghetti, P. Lanotte, A. Rosenau, L. W. van et al.** 2006. Prophagic DNA fragments in *Streptococcus agalactiae* strains and association with neonatal meningitis. J.Clin.Microbiol. **44**:1049-1058.

245. **Vergnaud, G. and F. Denoeud.** 2000. Minisatellites: Mutability and Genome Architecture. Genome Research **10**:899-907.

246. **Vogt, P.** 1990. Potential genetic functions of tandem repeated DNA sequence blocks in the human genome are based on a highly conserved "chromatin folding code". Hum.Genet. **84**:301-336.

247. **Wassenaar, T. M., J. A. Wagenaar, A. Rigter, C. Fearnley, D. G. Newell, and B. Duim.** 2002. Homonucleotide stretches in chromosomal DNA of *Campylobacter jejuni* display high frequency polymorphism as detected by direct PCR analysis. FEMS Microbiol.Lett. **212**:77-85.

248. **Wastfelt, M., M. Stalhammar-Carlemalm, A. M. Delisse, T. Cabezon, and G. Lindahl.** 1997. The Rib and alpha proteins define a family of group B streptococcal surface proteins that confer protective immunity. Adv.Exp.Med.Biol. **418**:619-622.

249. **Wayne, L. G., D. J. Brenner, R. R. Colwell, P. A. D. Grimont, O. Kandler, M. I. Krichevsky et al.** 1987. Report of the Ad Hoc Committee on Reconciliation of Approaches to Bacterial Systematics. International Journal of Systematic Bacteriology **37**:463-464.

250. **Weiser, J. N. and C. E. Rubens.** 1987. Transposon mutagenesis of group B streptococcus beta-hemolysin biosynthesis. Infect.Immun. **55**:2314-2316.

251. **Wen, L., Q. Wang, Y. Li, F. Kong, G. L. Gilbert, B. Cao et al.** 2006. Use of a serotype-specific DNA microarray for identification of group B Streptococcus (*Streptococcus agalactiae*). J.Clin.Microbiol. **44**:1447-1452.

252. **Wernecke, M., C. Mullen, V. Sharma, J. Morrison, T. Barry, M. Maher et al.** 2009. Evaluation of a novel real-time PCR test based on the *ssrA* gene for the identification of group B streptococci in vaginal swabs. BMC.Infect.Dis. **9**:148.

253. **Werner, G., I. Klare, and W. Witte.** 2007. The current MLVA typing scheme for *Enterococcus faecium* is less discriminatory than MLST and PFGE for epidemic-virulent, hospital-adapted clonal types. BMC Microbiology **7**:28.

254. **Wilgenbusch, J. C. and D. Swofford.** 2003. Inferring evolutionary trees with PAUP*. Curr.Protoc.Bioinformatics. **Chapter 6**:Unit.

255. **Williams, J. G., A. R. Kubelik, K. J. Livak, J. A. Rafalski, and S. V. Tingey.** 1990. DNA polymorphisms amplified by arbitrary primers are useful as genetic markers. Nucleic Acids Res. **18**:6531-6535.

256. **Wroblewski, D., G. E. Hannett, D. J. Bopp, G. K. Dumyati, T. A. Halse, N. B. Dumas et al.** 2009. Rapid molecular characterization of *Clostridium difficile* and assessment of populations of *C. difficile* in stool specimens. J.Clin.Microbiol. **47**:2142-2148.

257. **Yancey, M. K., T. Armer, P. Clark, and P. Duff.** 1992. Assessment of rapid identification tests for genital carriage of group B streptococci. Obstet.Gynecol. **80**:1038-1047.

258. **Yang, Z.** 1993. Maximum-likelihood estimation of phylogeny from DNA sequences when substitution rates differ over sites. Mol.Biol.Evol. **10**:1396-1401.

259. **Zaiss, N. H., M. Rupnik, E. J. Kuijper, C. Harmanus, D. Michielsen, K. Janssens et al.** 2009. Typing *Clostridium difficile* strains based on tandem repeat sequences. BMC.Microbiol. **9**:6.

260. **Zhu, Y., P. E. Fournier, H. Ogata, and D. Raoult.** 2005. Multispacer Typing of *Rickettsia prowazekii* Enabling Epidemiological Studies of Epidemic Typhus. Journal of Clinical Microbiology **43**:4708-4712.

261. **Zoysa, A., K. Edwards, S. Gharbia, A. Underwood, A. Charlett, and A. Efstratiou.** 2012. Non-culture detection of *Streptococcus agalactiae* (Lancefield group B Streptococcus) in clinical samples by real-time PCR. J.Med.Microbiol. **61**:1086-1090.

Chapter 9

Appendices

9.0 Appendices

9.1 Strain Collection

Strain	Serotype	Source	Type	Year	Isolation Site	Patient Age	Other Information
515	IA	ATCC	Reference		Not known	Not known	Neonate
9933	IA	ATCC	Reference		Not known	Not known	
6175	NT	NCTC	Reference	1941	Bovine	Not known	Bovine mastitis
8017	NT	NCTC	Reference	1949	Not known	Not known	Pigmented
8020	NT	NCTC	Reference	1949	Not known	Not known	Non-pigmented
8100	1c	NCTC	Reference	1949	Not known	Not known	
8181	1c	NCTC	Reference	1949	Milk	Not known	
8182	1d	NCTC	Reference	1949	Not known	Not known	
8183	2a	NCTC	Reference	1949	Not known	Not known	
8184	3a	NCTC	Reference	1949	Not known	Not known	
8186	3/C	NCTC	Reference	1950	Not known	Not known	
8188	6a	NCTC	Reference	1949	Not known	Not known	
8190	1a	NCTC	Reference	1949	Not known	Not known	
8541	NT	NCTC	Reference	1953	Human	Not known	Vaginal carrier
8542	NT	NCTC	Reference	1953	Human	Not known	Throat carrier
9409	NT	NCTC	Reference	1953	Human	Not known	Vaginal carrier
9410	NT	NCTC	Reference	1953	Human	Not known	Fatal infection
9411	NT	NCTC	Reference	1953	Human	Not known	Throat carrier
9412	NT	NCTC	Reference	1953	Human	Not known	Cervix, puerperal sepsis
9415	NT	NCTC	Reference	1953	Human	Not known	Nose
9828	R	NCTC	Reference	1956	Not known	Not known	
9829	R	NCTC	Reference	1956	Not known	Not known	
11080	X	NCTC	Reference	1976	Bovine	Not known	Chronic mastitis
11930	IV	NCTC	Reference	1985	Not known	Not known	
12906	NT	NCTC	Reference	2007	Not known	Not known	
12907	NT	NCTC	Reference	2007	Not known	Not known	
01/59	VII	HPA Reference Lab	Not known	2001	Not known	Not known	
01/08	VII	HPA Reference Lab	Clinical	2001	Blood	0 years	
01/13	VII	HPA Reference Lab	Clinical	2001	Placenta	23 years	
00/46	VIII	HPA Reference Lab	Clinical	2000	Blood	0 years	
00/465	VII	HPA Reference Lab	Clinical	2000	CSF	1 year	
01/776	VIII	HPA Reference Lab	Not known	2001	Not known	Not known	
01/789	VIII	HPA Reference Lab	Not known	2001	Not known	Not known	
03/0021	VII	HPA Reference Lab	Clinical	2003	Blood	69 years	
03/0476	III	HPA Reference Lab	Clinical	2003	Blood	1 month	
03/0477/GB	IA	HPA Reference Lab	Clinical	2003	Blood	66 years	
03/177	IV	HPA Reference Lab	Clinical	2003	Blood	37 years	
03/207	VI	HPA Reference Lab	Clinical	2003	Blood	74 years	
03/215	VI	HPA Reference Lab	Clinical	2003	Blood	42 years	
03/226	IV	HPA Reference Lab	Clinical	2003	Blood	0 years	
03/247	IV	HPA Reference Lab	Clinical	2003	Blood	83 years	
03/270	IV	HPA Reference Lab	Clinical	2003	Urine	17 years	
03/323	VII	HPA Reference Lab	Clinical		Not known	Not known	
03/414	IB	HPA Reference Lab	Clinical		Not known	Not known	
03/424	V	HPA Reference Lab	Clinical	2003	Blood	0 years	
03/427	IB	HPA Reference Lab	Clinical	2003	CSF	3 months	
03/431	V	HPA Reference Lab	Not known	2003	Not known	Not known	
03/438	V	HPA Reference Lab	Clinical	2003	Swab	0 years	
03/439	II	HPA Reference Lab	Clinical	2003	Not known	Not known	Post mortem
03/442	IB	HPA Reference Lab	Clinical	2003	Blood	54 years	
03/443	V	HPA Reference Lab	Clinical	2003	Blood	29 years	
03/447	II	HPA Reference Lab	Clinical	2003	Lung	2 months	Post mortem
03/451	IA	HPA Reference Lab	Clinical	2003	Blood	0 years	
03/453	IB	HPA Reference Lab	Clinical	2003	Blood	0 years	
03/460	1b/c	HPA Reference Lab	Clinical	2003	Blood	22 years	

03/463	IA	HPA Reference Lab	Clinical	2003	Blood	Not known	
03/464	III	HPA Reference Lab	Clinical	2003	Blood	0 years	
03/467	V	HPA Reference Lab	Clinical	2003	Blood	32 years	
03/469	II	HPA Reference Lab	Clinical	2003	Vaginal Swab	28 years	
03/470	II	HPA Reference Lab	Clinical	2003	Ear swab	0 years	
03/471	IA	HPA Reference Lab	Clinical	2003	Blood	49 years	
03/474	IV	HPA Reference Lab	Clinical	2003	Blood	0 years	
03/478/GB	III	HPA Reference Lab	Clinical	2003	Blood	0 years	
03/479/GB	III	HPA Reference Lab	Clinical	2003	Blood	74 years	
03-0475	III	HPA Reference Lab	Clinical	2003	Blood	0 years	
18RS21	II	NCTC	Reference	1976	Human	Not known	Throat
2603V/R	V	ATCC	Reference	2003	Human	Not known	
A909	IA	NCTC	Reference	1976	Not known	Not known	
CJB111	V	ATCC	Reference	1990	Human	Not known	Blood
COH1	III	ATCC	Reference	1985	Human	Not known	Blood
GBS X	III	HPA Reference Lab	Reference		Not known	Not known	
GBS-C	C	HPA Reference Lab	Reference		Not known	Not known	
GBS-IA	IA	HPA Reference Lab	Reference		Not known	Not known	
GBS-IB	IB	HPA Reference Lab	Reference		Not known	Not known	
GBS-II	II	HPA Reference Lab	Reference		Not known	Not known	
GBS-III	III	HPA Reference Lab	Reference		Not known	Not known	
GBS-IV	IV	HPA Reference Lab	Reference		Not known	Not known	
GBS-NT-03/454/GB	NT	HPA Reference Lab	Reference		Not known	Not known	
GBS-NT-03/455/GB	NT	HPA Reference Lab	Reference		Not known	Not known	
GBS-R	R	HPA Reference Lab	Reference		Not known	Not known	
GBS-Type	NT	HPA Reference Lab	Reference		Not known	Not known	
GBS-V	V	HPA Reference Lab	Reference		Not known	Not known	
GBS-VI	VI	HPA Reference Lab	Reference		Not known	Not known	
GBS-VII	VII	HPA Reference Lab	Reference		Not known	Not known	
GBS-VIII	VIII	HPA Reference Lab	Reference		Not known	Not known	
H034520043	III	HPA Reference Lab	Clinical	2003	Blood	25 days	
H034540199	IA/C	HPA Reference Lab	Clinical	2003	Blood	68 years	
H034540202	IB/C	HPA Reference Lab	Clinical	2003	Groin swab	0 days	
H034580222	II	HPA Reference Lab	Clinical	2003	Lung/spleen	8 days	
H034600053	II/C	HPA Reference Lab	Clinical	2003	Blood	0 days	
H034620029	V	HPA Reference Lab	Clinical	2003	Blood	7 days	Post mortem
H034640397	IB/C	HPA Reference Lab	Clinical	2003	Lung/spleen	11 days	
H034700302	IA/C	HPA Reference Lab	Clinical	2003	Ear swab	4 days	
H034760020	V	HPA Reference Lab	Clinical	2003	Blood	87 years	
H034780226	IA	HPA Reference Lab	Clinical	2003	Blood	31 years	Post natal
H034940026	III/X	HPA Reference Lab	Clinical	2003	CSF + blood	24 days	
H034960218	VI/C	HPA Reference Lab	Clinical	2003	Blood	63 years	
H034980044	NT	HPA Reference Lab	Clinical	2003	Elbow aspirate	57 years	
H035140030	IB	HPA Reference Lab	Clinical	2003	Blood	3 months	
H040120318	IA	HPA Reference Lab	Clinical	2004	Blood	89 years	
H040200291	II	HPA Reference Lab	Clinical	2004	Blood	45 years	
H040420423	IV/C	HPA Reference Lab	Clinical	2004	Breast milk	29 years	Post natal
H040440330	II/C	HPA Reference Lab	Clinical	2004	Joint aspirate	81 years	
H040540417	II/C	HPA Reference Lab	Clinical	2004	Blood	36 years	Maternity
H040600213	NT	HPA Reference Lab	Clinical	2004	Blood	69 years	
H040680236	IV	HPA Reference Lab	Clinical	2004	Blood	4 days	
H040840193	V	HPA Reference Lab	Clinical	2004	Blood	74 years	
H040880024	IV/C	HPA Reference Lab	Clinical	2004	Nose swab	18 days	Post mortem
H041060026	II/C	HPA Reference Lab	Clinical	2004	Blood	29 years	
H041100399	III/X	HPA Reference Lab	Clinical	2004	Blood	2 days	
H041180103	II	HPA Reference Lab	Clinical	2004	Blood	25 years	
H041200156	III	HPA Reference Lab	Clinical	2004	CSF	5 days	
H041360432	IV/C	HPA Reference Lab	Clinical	2004	Blood	0 years	
H041420037	II	HPA Reference Lab	Clinical	2004	CSF	11 years	
H041420425	IB/C	HPA Reference Lab	Clinical	2004	Blood	64 years	
H041540478	IA	HPA Reference Lab	Clinical	2004	Blood	77 years	
H041740496	II	HPA Reference Lab	Clinical	2004	Blood	78 years	
H041800048	IV	HPA Reference Lab	Clinical	2004	Blood	8 days	
H041860027	V	HPA Reference Lab	Clinical	2004	Blood	Not known	

H041960046	IB	HPA Reference Lab	Clinical	2004	Blood	Not known	
H041960680	III	HPA Reference Lab	Clinical	2004	Blood	8 days	
H042380454	III	HPA Reference Lab	Clinical	2004	Blood	0 years	
H042560014	IB/C	HPA Reference Lab	Clinical	2004	Blood	3 days	
H042680168	V	HPA Reference Lab	Clinical	2004	Blood	0 years	
H042700323	V	HPA Reference Lab	Clinical	2004	Blood	6 days	
H042760726	IB/C	HPA Reference Lab	Clinical	2004	Blood	60 years	
H042880027	IA	HPA Reference Lab	Clinical	2004	Blood	96 years	
H042960014	IA	HPA Reference Lab	Clinical	2004	Blood	1 month	
H043220036	IA	HPA Reference Lab	Clinical	2004	Blood	41 years	Maternity
H043260048	II	HPA Reference Lab	Clinical	2004	Blood	0 years	Post mortem
H043340011	III	HPA Reference Lab	Clinical	2004	Blood	74 years	
H043340012	IA/C	HPA Reference Lab	Clinical	2004	Blood	41 years	
H36B	IB	NCTC	Reference	1949	Not known	N/a	
NEM316	III	NCTC	Reference		Not known	N/a	Fatal septicemia

9.2 MLST Allelic Profiles

Strain	*adhP*	*pheS*	*atr*	*glnA*	*sdhA*	*glcK*	tkt	ST	Serotype
01/59	1	1	2	1	1	2	2	1	VII
00/46	1	1	2	1	1	2	2	1	VIII
00/465	1	1	2	1	1	2	2	1	VII
03/431	1	1	2	1	1	2	2	1	V
03/467	1	1	2	1	1	2	2	1	V
CJB111	1	1	2	1	1	2	2	1	V
GBS-VII	1	1	2	1	1	2	2	1	VII
GBS-VIII	1	1	2	1	1	2	2	1	VIII
H041060026	1	1	2	1	1	2	2	1	II/C
03/177	1	1	3	1	1	2	2	2	IV
03/247	1	1	3	1	1	2	2	2	IV
03/469	1	1	3	1	1	2	2	2	II
H040440330	1	1	3	1	1	2	2	2	II/C
03/470	1	1	3	1	1	2	2	2	II
GBS-NT-03/455/GB	1	1	3	1	1	2	2	2	NT
8542	9	1	2	1	3	2	2	6	Unknown
9412	9	1	2	1	3	2	2	6	Unknown
9415	9	1	2	1	3	2	2	6	Unknown
H36B	9	1	2	1	3	2	2	6	IB
GBS-IB	9	1	2	1	3	2	2	6	IB
12907	10	1	2	1	3	2	2	7	Unknown
03/471	10	1	2	1	3	2	2	7	IA
A909	10	1	2	1	3	2	2	7	IA
GBS-C	10	1	2	1	3	2	2	7	C
H041100399	10	1	2	1	3	2	2	7	III/X
03/427	4	1	4	1	3	3	2	8	IB
03/453	4	1	4	1	3	3	2	8	IB
H035140030	4	1	4	1	3	3	2	8	IB
H042560014	4	1	4	1	3	3	2	8	IB/C
H041180103	4	1	4	1	3	3	2	8	II
03/460	8	1	4	1	3	3	2	9	Unknown
H042760726	8	1	4	1	3	3	2	9	IB/C
11930	9	1	4	1	3	3	2	10	IV
03/442	9	1	4	1	3	3	2	10	IB
GBS-IV	9	1	4	1	3	3	2	10	IV
H034540202	9	1	4	1	3	3	2	10	IB/C
H041960046	9	1	4	1	3	3	2	10	IB
H040880024	10	1	4	1	3	3	2	12	IV/C
H040600213	10	1	4	1	3	3	2	12	NT
H041420425	10	1	4	1	3	3	2	12	IB/C
GBS-VI	1	1	2	1	5	2	2	14	VI
03/464	2	1	1	2	1	1	1	17	III
COH1	2	1	1	2	1	1	1	17	III
H03420043	2	1	1	2	1	1	1	17	III
H034620029	2	1	1	2	1	1	1	17	V
H034940026	2	1	1	2	1	1	1	17	III/X
H040420423	2	1	1	2	1	1	1	17	IV/C
H041360432	2	1	1	2	1	1	1	17	IV/C
H041960680	2	1	1	2	1	1	1	17	III
H042380454	2	1	1	2	1	1	1	17	III
8541	1	1	3	2	2	2	2	19	Unknown
18RS21	1	1	3	2	2	2	2	19	II
GBS-II	1	1	3	2	2	2	2	19	II

03/478/GB	1	1	3	2	2	2	2	19	III
H040200291	1	1	3	2	2	2	2	19	II
H034600053	13	3	1	3	1	1	1	22	II/C
H034960218	13	3	1	3	1	1	1	22	VI/C
H040540417	13	3	1	3	1	1	1	22	II/C
H041420037	13	3	1	3	1	1	1	22	II
8186	5	4	6	3	2	1	3	23	3/C
9409	5	4	6	3	2	1	3	23	Unknown
9410	5	4	6	3	2	1	3	23	Unknown
9411	5	4	6	3	2	1	3	23	Unknown
03/0477/GB	5	4	6	3	2	1	3	23	IA
03/451	5	4	6	3	2	1	3	23	IA
03/463	5	4	6	3	2	1	3	23	IA
H034700302	5	4	6	3	2	1	3	23	IA/C
H041540478	5	4	6	3	2	1	3	23	IA
H042960014	5	4	6	3	2	1	3	23	IA
H043220036	5	4	6	3	2	1	3	23	IA
H043340012	5	4	6	3	2	1	3	23	IA/C
NEM316	5	4	6	3	2	1	3	23	III
515	5	4	6	3	2	1	3	23	IA
12906	5	4	6	3	2	1	3	23	Unknown
H034540119	5	4	4	3	2	3	3	24	IA/C
H042880027	5	4	4	3	2	3	3	24	IA
0933	5	4	6	3	8	1	3	25	IA
GBS-IA	5	4	6	3	8	1	3	25	IA
GBS-Type	5	4	6	3	8	1	3	25	Unknown
GBS-V	1	1	5	4	1	4	6	26	V
03/439	1	1	3	5	2	2	2	28	II
H034580222	1	1	3	5	2	2	2	28	II
H040120318	1	1	3	5	2	2	2	28	IA
03/447	1	1	3	5	2	2	2	28	II
H043260048	1	1	3	5	2	2	2	28	II
03/270	1	1	3	2	1	2	2	44	IV
03/479/GB	1	1	3	2	1	2	2	44	III
03/0476	2	1	1	2	2	1	1	48	III
03/414	10	1	3	1	3	2	2	51	IB
8100	13	1	1	13	1	1	1	61	1c
8181	13	1	1	13	1	1	1	61	1c
8182	13	1	1	13	1	1	1	61	1d
8188	13	1	1	13	1	1	1	61	6a
9828	13	1	1	13	1	1	5	67	R
GBS-R	13	1	1	13	1	1	5	67	R
2603VR	1	1	3	2	2	2	9	110	V
8017	5	4	1	3	2	1	3	144	Unknown
8020	5	4	1	3	2	1	3	144	Unknown
H034780226	1	1	3	1	1	3	4	172	IA
6175	2	1	1	3	1	1	1	174	Unknown
8184	2	1	1	3	1	1	1	174	3a
03-0475	1	1	3	2	18	2	2	182	III

9.3 Three gene Allelic Profiles

Isolate	*cpsL*	SAG0043	SAG1894	ST
8182	1	7	12	1
11930	2	3	4	2
NEM316	3	4	5	3
12907	4	3	4	4
8542	5	3	4	5
8017	6	4	5	6
01/59	7	1	4	7
01/08	7	1	4	7
01/13	7	1	4	7
03-323	7	1	4	7
GBS-VII	7	1	4	7
00-46	7	12	4	8
01-789	8	1	4	9
GBS-VIII	8	1	4	9
00-465	8	12	4	10
GBS-VI	9	1	4	11
01-776	9	2	4	12
03-0021	9	3	4	13
03-215	9	3	4	13
A909	10	3	4	14
12906	10	4	3	15
515	10	4	5	16
993	10	4	5	16
8020	10	4	5	16
8186	10	4	5	16
9409	10	4	5	16
9410	10	4	5	16
9411	10	4	5	16
03-0477-GB	10	4	5	16
03-451	10	4	5	16
03-463	10	4	5	16
GBS-IA	10	4	5	16
GBS-Type	10	4	5	16
H034540119	10	4	5	16
H034700302	10	4	5	16
H041540478	10	4	5	16
H042880027	10	4	5	16
H043220036	10	4	5	16
H042960014	10	17	5	17
H043340012	10	17	5	17
03-177	11	1	4	18
03-226	11	1	4	18
03-270	11	1	4	18

03-474	11	1	4	18
H040440330	11	1	4	18
H041200156	11	1	4	18
H041800048	11	1	4	18
03-443	11	1	5	19
03-247	11	1	6	20
GBS-IV	11	3	4	21
H040840193	11	3	4	21
03-207	12	1	4	22
9412	13	3	4	23
9415	13	3	4	23
03-414	13	3	4	23
03-427	13	3	4	23
03-453	13	3	4	23
GBS-IB	13	3	4	23
H0345020	13	3	4	23
H041180103	13	3	4	23
H041960046	13	3	4	23
H042560014	13	3	4	23
H36B	13	3	4	23
H04276-0726	13	3	14	24
03-460	13	6	4	25
H04142-0425	13	16	2	26
03-424	14	1	4	27
03-431	14	1	4	27
03-467	14	1	4	27
CJB111	14	1	4	27
GBS-NT-03-455-GB	14	1	4	27
H034980044	14	1	4	27
H040680236	14	1	4	27
H041060026	14	1	4	27
H041860077	14	1	4	27
H042680168	14	1	4	27
H042700325	14	1	4	27
2603VR	14	1	5	28
H041100399	14	3	4	29
GBS-V	14	11	4	30
H034760020	14	14	4	31
03-469	15	1	1	32
03-470	15	1	4	33
H034640397	15	1	4	33
H041740496	15	1	4	33
8541	15	1	5	34
03-439	15	1	5	34
03-447	15	1	5	34
18RS21	15	1	5	34

GBS-II	15	1	5	34
GBS-NT-03-454-GB	15	1	5	34
H040120318	15	1	5	34
H040200291	15	1	5	34
H043260048	15	1	5	34
03-0476-GB	15	1	7	35
03-438	15	1	IS1381	36
H040600213	15	3	2	37
H040880024	15	3	2	37
8181	15	7	9	38
H034600053	15	7	13	39
H034960218	15	7	13	39
H040540417	15	7	13	39
8100	15	9	10	40
H034580222	15	12	5	41
GBS-X	15	13	10	42
H04142-0037	15	15	13	43
03-442	16	3	4	44
03-464	17	5	8	45
COH1	17	5	8	45
H034620029	17	5	8	45
H034940026	17	5	8	45
H041360432	17	5	8	45
H041960680	17	5	8	45
H042380454	17	5	8	45
03-471	18	3	4	46
03-478-GB	19	1	5	47
03-479-GB	19	1	5	47
H04334-0011	19	1	5	47
11080	20	7	9	48
GBS-III	20	7	9	48
8184	21	5	5	49
6175	21	8	5	50
03-0475	22	7	11	51
H0347800226	23	1	4	52
8183	23	10	4	53
8188	25	7	10	54
8190	26	1	4	55
9828	26	7	4	56
GBS-R	26	7	4	56
H03420043	27	5	8	57
H035140030	28	3	4	58
H040420423	29	5	8	59
9829	30	9	10	60
GBS-C	31	1	4	61

9.4 Four Gene Allelic Profiles

Isolate	*cpsL*	SAG0043	SAG1894	*valS*	ST
8182	1	7	12	9	1
11930	2	3	4	2	2
NEM316	3	4	5	8	3
12907	4	3	4	2	4
8542	5	3	4	2	5
8017	6	4	5	8	6
01/59	7	1	4	1	7
01/08	7	1	4	1	7
01/13	7	1	4	1	7
03-323	7	1	4	1	7
GBS-VII	7	1	4	1	7
00-46	7	12	4	1	8
01-789	8	1	4	1	9
GBS-VIII	8	1	4	1	9
00-465	8	12	4	1	10
GBS-VI	9	1	4	4	11
01-776	9	2	4	9	12
03-0021	9	3	4	2	13
03-215	9	3	4	5	14
A909	10	3	4	2	15
12906	10	4	3	3	16
H034540119	10	4	5	2	17
H042880027	10	4	5	2	17
515	10	4	5	3	18
03-0477-GB	10	4	5	3	18
03-451	10	4	5	3	18
03-463	10	4	5	3	18
H034700302	10	4	5	3	18
H041540478	10	4	5	3	18
H043220036	10	4	5	3	18
993	10	4	5	8	19
8020	10	4	5	8	19
8186	10	4	5	8	19
9409	10	4	5	8	19
9410	10	4	5	8	19
9411	10	4	5	8	19
GBS-IA	10	4	5	8	19
GBS-Type	10	4	5	8	19
H042960014	10	17	5	3	20
H043340012	10	17	5	3	20
03-177	11	1	4	4	21
03-226	11	1	4	4	21
03-270	11	1	4	4	21

H040440330	11	1	4	4	21
H041200156	11	1	4	4	21
H041800048	11	1	4	4	21
03-474	11	1	4	6	22
03-443	11	1	5	4	23
03-247	11	1	6	4	24
GBS-IV	11	3	4	2	25
H040840193	11	3	4	2	25
03-207	12	1	4	1	26
9412	13	3	4	2	27
9415	13	3	4	2	27
03-414	13	3	4	2	27
03-427	13	3	4	2	27
03-453	13	3	4	2	27
GBS-IB	13	3	4	2	27
H0345020	13	3	4	2	27
H041180103	13	3	4	2	27
H041960046	13	3	4	2	27
H042560014	13	3	4	2	27
H36B	13	3	4	2	27
H042760726	13	3	14	2	28
03-460	13	6	4	2	29
H041420425	13	16	2	2	30
03-424	14	1	4	1	31
03-431	14	1	4	1	31
03-467	14	1	4	1	31
CJB111	14	1	4	1	31
GBS-NT-03-455-GB	14	1	4	1	31
H034980044	14	1	4	1	31
H040680236	14	1	4	1	31
H041060026	14	1	4	1	31
H041860077	14	1	4	1	31
H042680168	14	1	4	1	31
H042700325	14	1	4	1	31
2603VR	14	1	5	4	32
H041100399	14	3	4	8	33
GBS-V	14	11	4	14	34
H034760020	14	14	4	1	35
03-469	15	1	1	4	36
H034640397	15	1	4	1	37
03-470	15	1	4	4	38
H041740496	15	1	4	4	38
8541	15	1	5	4	39
03-439	15	1	5	4	39
03-447	15	1	5	4	39
18RS21	15	1	5	4	39

GBS-II	15	1	5	4	39
GBS-NT-03-454-GB	15	1	5	4	39
H040120318	15	1	5	4	39
H040200291	15	1	5	4	39
H043260048	15	1	5	16	40
03-0476-GB	15	1	7	2	41
03-438	15	1	IS1381	4	42
H040600213	15	3	2	2	43
H040880024	15	3	2	2	43
8181	15	7	9	9	44
H034600053	15	7	13	15	45
H034960218	15	7	13	15	45
H040540417	15	7	13	15	45
8100	15	9	10	9	46
H034580222	15	12	5	4	47
GBS-X	15	13	10	9	48
H041420037	15	15	13	15	49
03-442	16	3	4	2	50
03-464	17	5	8	4	51
COH1	17	5	8	6	52
H034620029	17	5	8	6	52
H034940026	17	5	8	6	52
H041360432	17	5	8	6	52
H041960680	17	5	8	6	52
H042380454	17	5	8	6	52
03-471	18	3	4	7	53
03-478-GB	19	1	5	4	54
H043340011	19	1	5	4	54
03-479-GB	19	1	5	8	55
11080	20	7	9	9	56
GBS-III	20	7	9	9	56
8184	21	5	5	10	57
6175	21	8	5	10	58
03-0475	22	7	11	4	59
H0347800226	23	1	4	8	60
8183	23	10	4	11	61
8188	25	7	10	9	62
8190	26	1	4	3	63
9828	26	7	4	13	64
GBS-R	26	7	4	13	64
H03420043	27	5	8	6	65
H035140030	28	3	4	2	66
H040420423	29	5	8	6	67
9829	30	9	10	9	68

GBS-C	31	1	4	4	69

9.5 MNR Repeat Containing Non coding Region Allelic Profiles

Isolate	SAG0032	SAG0649	SAG1768	SAK_1320	ST
03-437	1	1	1	GBSi1	1
01/59	1	1	2	1	2
01/08	1	1	2	1	2
01/13	1	1	2	1	2
00-46	1	1	2	1	2
00-465	1	1	2	1	2
01-789	1	1	2	1	2
03-207	1	1	2	1	2
03-323	1	1	2	1	2
03-424	1	1	2	1	2
03-431	1	1	2	1	2
03-443	1	1	2	1	2
03-467	1	1	2	1	2
CJB111	1	1	2	1	2
GBS-NT-03-455-GB	1	1	2	1	2
GBS-VI	1	1	2	1	2
GBS-VII	1	1	2	1	2
GBS-VIII	1	1	2	1	2
H034760020	1	1	2	1	2
H0347800226	1	1	2	1	2
H034980044	1	1	2	1	2
H040680236	1	1	2	1	2
H041060026	1	1	2	1	2
H041200156	1	1	2	1	2
H041800048	1	1	2	1	2
H041860077	1	1	2	1	2
H042700325	1	1	2	1	2
H042680168	1	1	2	5	3
03-469	1	1	2	GBSi1	4
03-470	1	1	2	GBSi1	4
H034640397	1	1	2	GBSi1	4
8183	1	1	4	1	5
8190	1	1	4	1	5
8541	1	1	5	1	6
GBS-II	1	1	5	1	6
H040120318	1	1	5	1	6
03-439	1	1	5	GBSi1	7
03-447	1	1	5	GBSi1	7
18RS21	1	1	5	GBSi1	7
2603VR	1	1	5	GBSi1	7
GBS-NT-03-454-GB	1	1	5	GBSi1	7
H034580222	1	1	5	GBSi1	7
H043260048	1	1	5	GBSi1	7

03-478-GB	1	1	5	IS1548	8
H043340011	1	1	5	IS1548	8
03-479-GB	1	1	14	IS1548	9
03-177	1	2	2	1	10
03-226	1	2	2	1	10
03-247	1	2	2	1	10
03-270	1	2	2	1	10
03-464	1	2	2	1	10
9409	1	2	4	1	11
9410	1	2	4	1	11
9411	1	2	4	1	11
NEM316	1	2	4	1	11
8017	1	2	4	7	12
8020	1	2	4	7	12
GBS-C	1	2	9	1	13
H041740496	1	NA	2	1	NA
11930	2	1	2	1	14
12907	2	1	2	1	14
03-414	2	1	2	1	14
03-442	2	1	2	1	14
03-453	2	1	2	1	14
03-471	2	1	2	1	14
A909	2	1	2	1	14
GBS-IV	2	1	2	1	14
H0345020	2	1	2	1	14
H035140030	2	1	2	1	14
H040600213	2	1	2	1	14
H040840193	2	1	2	1	14
H040880024	2	1	2	1	14
H041180103	2	1	2	1	14
H041420425	2	1	2	1	14
H041960046	2	1	2	1	14
H042760726	2	1	2	1	14
01-776	2	1	2	2	15
03-215	2	1	2	2	15
03-460	2	1	2	3	16
H042560014	2	1	2	4	17
GBS-IB	2	1	2	7	18
03-0021	2	1	3	2	19
03-427	2	1	13	1	20
9412	2	2	2	1	21
8542	2	2	2	7	22
9415	2	2	2	7	22
H36B	2	2	2	7	22
H041100399	2	3	2	1	23
515	3	1	4	1	24
8100	3	1	4	1	24

11080	3	1	4	1	24
03-0477-GB	3	1	4	1	24
03-451	3	1	4	1	24
GBS-III	3	1	4	1	24
GBS-R	3	1	4	1	24
H034540119	3	1	4	1	24
H034700302	3	1	4	1	24
H041540478	3	1	4	1	24
H042880027	3	1	4	1	24
H042960014	3	1	4	1	24
H043220036	3	1	4	1	24
H043340012	3	1	4	1	24
8181	3	1	7	1	25
GBS-X	3	1	GBSi1	1	26
9829	3	2	2	1	27
993	3	2	4	1	28
8186	3	2	4	1	28
9828	3	2	4	1	28
03-463	3	2	4	1	28
GBS-IA	3	2	4	1	28
GBS-Type	3	2	4	1	28
12906	3	2	4	6	29
8188	3	2	8	1	30
03-0475	3	2	ISSag8	1	31
8182	3	2	*S. equi* transposase	1	32
03-0476-GB	4	1	5	1	33
H040200291	4	NA	5	GBSi1	NA
H03420043	5	1	5	1	34
H041360432	5	1	5	1	34
03-474	5	1	5	GBSi1	35
COH1	5	1	5	GBSi1	35
H034620029	5	1	5	GBSi1	35
H034940026	5	1	5	GBSi1	35
H040420423	5	1	5	GBSi1	35
H041960680	5	1	5	GBSi1	35
H042380454	5	1	5	GBSi1	35
6175	5	1	6	1	36
8184	5	1	6	1	36
GBS-V	6	1	10	1	37
H040440330	7	2	2	1	38
H034600053	8	1	11	IS1548	39
H034960218	8	1	11	IS1548	39
H040540417	8	1	11	IS1548	39
H041420037	8	1	11	IS1548	39

9.6 Non coding Regions without MNR repeat Allelic Profiles

Isolate	SAG0043	SAG0106	SAG2143	ST
03-0021	1	2	2	1
03-215	1	2	2	1
8542	1	2	5	2
9412	1	2	5	2
9415	1	2	5	2
11930	1	2	5	2
12907	1	2	5	2
03-427	1	2	5	2
03-442	1	2	5	2
03-453	1	2	5	2
03-460	1	2	5	2
03-471	1	2	5	2
A909	1	2	5	2
GBS_IB	1	2	5	2
GBS_IV	1	2	5	2
H0345020	1	2	5	2
H035140030	1	2	5	2
H040600213	1	2	5	2
H040840193	1	2	5	2
H040880024	1	2	5	2
H041100399	1	2	5	2
H041420425	1	2	5	2
H041960046	1	2	5	2
H042560014	1	2	5	2
H042760726	1	2	5	2
H36B	1	2	5	2
03-414	2	2	5	3
6175	3	1	3	4
8184	3	1	3	4
H034620029	3	1	3	4
H034940026	3	1	3	4
H041360432	3	1	3	4
03-474	3	4	3	5
COH1	3	4	3	5
H03420043	3	4	3	5
H040420423	3	4	3	5
H041960680	3	4	3	5
H042380454	3	4	3	5
9828	4	1	1	6
GBS_R	4	1	1	6
8100	4	1	7	7
8181	4	1	7	7
8182	4	1	7	7
8188	4	1	7	7
9829	4	1	7	7

11080	4	1	7	7
03-0475	4	1	7	7
GBS_III	4	1	7	7
GBS_X	4	1	7	7
H034600053	4	1	7	7
H034960218	4	1	7	7
H040540417	4	1	7	7
H041420037	4	1	7	7
H041180103	5	2	5	8
8183	6	1	1	9
01/59	6	1	1	9
01/08	6	1	1	9
01/13	6	1	1	9
00-46	6	1	1	9
00-465	6	1	1	9
01-789	6	1	1	9
03-177	6	1	1	9
03-207	6	1	1	9
03-247	6	1	1	9
03-270	6	1	1	9
03-323	6	1	1	9
03-424	6	1	1	9
03-431	6	1	1	9
03-437	6	1	1	9
03-443	6	1	1	9
03-464	6	1	1	9
03-467	6	1	1	9
CJB111	6	1	1	9
GBS_C	6	1	1	9
GBS_NT_03-455-GB	6	1	1	9
GBS_VII	6	1	1	9
GBS_VIII	6	1	1	9
H034640397	6	1	1	9
H034760020	6	1	1	9
H034980044	6	1	1	9
H040440330	6	1	1	9
H040680236	6	1	1	9
H041060026	6	1	1	9
H041200156	6	1	1	9
H041800048	6	1	1	9
H041860077	6	1	1	9
H042680168	6	1	1	9
H042700325	6	1	1	9
8541	6	1	3	10
03-447	6	1	3	10
03-478-GB	6	1	3	10
03-479-GB	6	1	3	10
2603VR	6	1	3	10

H043260048	6	1	3	10
H043340011	6	1	3	10
03-226	6	1	4	11
03-469	6	1	6	12
03-470	6	1	6	12
18RS21	6	1	8	13
GBS-II	6	1	8	13
GBS-VI	6	2	3	14
03-0476-GB	6	3	3	15
H04020-0291	6	3	3	15
8190	6	6	1	16
03-0477-GB	7	2	3	17
03-439	8	1	3	18
515	9	2	3	19
993	9	2	3	19
8186	9	2	3	19
9409	9	2	3	19
9410	9	2	3	19
9411	9	2	3	19
12906	9	2	3	19
03-451	9	2	3	19
03-463	9	2	3	19
GBS_IA	9	2	3	19
GBS_Type	9	2	3	19
H03454-0119	9	2	3	19
H03470-0302	9	2	3	19
H04154-0478	9	2	3	19
H04288-0027	9	2	3	19
H04296-0014	9	2	3	19
H04322-0036	9	2	3	19
H04334-0012	9	2	3	19
NEM316	9	2	3	19
8017	9	5	3	20
8020	9	5	3	20
GBS-NT-03-454-GB	10	1	3	21
H034580222	11	1	3	22
H040120318	11	1	3	22
H0347800226	12	1	1	23
H041740496	NA	1	1	NA
GBS-V	NA	1	2	NA
01-776	NA	2	3	NA

9.7 Allele Typing Results for the Three Gene plus Insertion Sequence Typing

Isolate	*cpsL*	SAG0043	SAG1894	SAK_1320	ST
8182	1	7	12	1	1
11930	2	3	4	1	2
NEM316	3	4	5	1	3
12907	4	3	4	1	4
8542	5	3	4	7	5
8017	6	4	5	7	6
21551	7	1	4	1	7
40756	7	1	4	1	7
41275	7	1	4	1	7
03-323	7	1	4	1	7
GBS-VII	7	1	4	1	7
00-46	7	12	4	1	8
01-789	8	1	4	1	9
GBS-VIII	8	1	4	1	9
00-465	8	12	4	1	10
GBS-VI	9	1	4	1	11
01-776	9	2	4	2	12
03-0021	9	3	4	2	13
03-215	9	3	4	2	13
A909	10	3	4	1	14
12906	10	4	3	6	15
515	10	4	5	1	16
993	10	4	5	1	16
8186	10	4	5	1	16
9409	10	4	5	1	16
9410	10	4	5	1	16
9411	10	4	5	1	16
03-0477-GB	10	4	5	1	16
03-451	10	4	5	1	16
03-463	10	4	5	1	16
GBS-IA	10	4	5	1	16
GBS-Type	10	4	5	1	16
H034540119	10	4	5	1	16
H034700302	10	4	5	1	16
H041540478	10	4	5	1	16
H042880027	10	4	5	1	16
H043220036	10	4	5	1	16
8020	10	4	5	7	17
H042960014	10	17	5	1	18
H043340012	10	17	5	1	18
03-177	11	1	4	1	19
03-226	11	1	4	1	19
03-270	11	1	4	1	19
H040440330	11	1	4	1	19
H041200156	11	1	4	1	19
H041800048	11	1	4	1	19
03-474	11	1	4	GBSi1	20
03-443	11	1	5	1	21
03-247	11	1	6	1	22
GBS-IV	11	3	4	1	23

H040840193	11	3	4	1	23
03-207	12	1	4	1	24
9412	13	3	4	1	25
03-414	13	3	4	1	25
03-427	13	3	4	1	25
03-453	13	3	4	1	25
H0345020	13	3	4	1	25
H041180103	13	3	4	1	25
H041960046	13	3	4	1	25
H042560014	13	3	4	4	26
9415	13	3	4	7	27
GBS-IB	13	3	4	7	27
H36B	13	3	4	7	27
H04276-0726	13	3	14	1	28
03-460	13	6	4	3	29
H04142-0425	13	16	2	1	30
03-424	14	1	4	1	31
03-431	14	1	4	1	31
03-467	14	1	4	1	31
CJB111	14	1	4	1	31
GBS-NT-03-455-GB	14	1	4	1	31
H034980044	14	1	4	1	31
H040680236	14	1	4	1	31
H041060026	14	1	4	1	31
H041860077	14	1	4	1	31
H042700325	14	1	4	1	31
H042680168	14	1	4	5	32
2603VR	14	1	5	GBSi1	33
H041100399	14	3	4	1	34
GBS-V	14	11	4	1	35
H034760020	14	14	4	1	36
03-469	15	1	1	GBSi1	37
H041740496	15	1	4	1	38
03-470	15	1	4	GBSi1	39
H034640397	15	1	4	GBSi1	39
8541	15	1	5	1	40
GBS-II	15	1	5	1	40
H040120318	15	1	5	1	40
03-439	15	1	5	GBSi1	41
03-447	15	1	5	GBSi1	41
18RS21	15	1	5	GBSi1	41
GBS-NT-03-454-GB	15	1	5	GBSi1	41
H040200291	15	1	5	GBSi1	41
H043260048	15	1	5	GBSi1	41
03-0476-GB	15	1	7	1	42
03-438	15	1	IS1381	GBSi1	43
H040600213	15	3	2	1	44
H040880024	15	3	2	1	44
8181	15	7	9	1	45
H034600053	15	7	13	IS1548	46
H034960218	15	7	13	IS1548	46
H040540417	15	7	13	IS1548	46
8100	15	9	10	1	47
H034580222	15	12	5	GBSi1	48
GBS-X	15	13	10	1	49

H04142-0037	15	15	13	IS1548	50
03-442	16	3	4	1	51
03-464	17	5	8	1	52
H041360432	17	5	8	1	52
COH1	17	5	8	GBSi1	53
H034620029	17	5	8	GBSi1	53
H034940026	17	5	8	GBSi1	53
H041960680	17	5	8	GBSi1	53
H042380454	17	5	8	GBSi1	53
03-471	18	3	4	1	54
03-478-GB	19	1	5	IS1548	55
03-479-GB	19	1	5	IS1548	55
H04334-0011	19	1	5	IS1548	55
11080	20	7	9	1	56
GBS-III	20	7	9	1	56
8184	21	5	5	1	57
6175	21	8	5	1	58
03-0475	22	7	11	1	59
H0347800226	23	1	4	1	60
8183	23	10	4	1	61
8188	25	7	10	1	62
8190	26	1	4	1	63
9828	26	7	4	1	64
GBS-R	26	7	4	1	64
H03420043	27	5	8	1	65
H035140030	28	3	4	1	66
H040420423	29	5	8	GBSi1	67
9829	30	9	10	1	68
GBS-C	31	1	4	1	69

9.8 Core Genome Scripts

9.8.1 Gene Extraction Script

```
#!/usr/bin/perl
#
#
#
#geneextraction.pl
#Removes CDS fom genbank file


#use lib "/usr/local/share/perl/5.8.8/Bio";
use Bio::SeqIO;
use Bio::AnnotationCollectionI;
chdir "/home/richard/Sequences_2/";
@filename= <*.gb>;
#@filename=("B_halodurans.gbk",
"B_licheniformis_AE017333.gbk","B_licheniformis_CP000002.gbk","B_thuringiensis.gbk");
#@filename=("B_thuringiensis.gbk");
foreach $filename(@filename){
    my $seqio=Bio::SeqIO->new('-format'=>'GenBank',-file=> $filename);
    my $seq=$seqio->next_seq;
    @feats = $seq->get_SeqFeatures;
    foreach $feature(@feats) {
        $tag = $feature->primary_tag;
        if ($tag=~/CDS/){
            @tagvalues = $feature->get_tag_values ('locus_tag');
            $locus_name="";
            foreach $tagvalue(@tagvalues){
                $locus_name = $locus_name . $tagvalue;
                print "$tagvalue\n";
            }
            $locus_name=~s/\///g; # this will remove any extra slashes (forward) that
            might be present in the locus name as this will throwthe script off track as
            cant process this

            $location = $feature->location;
            $start = $location->start;
            $end = $location->end;
            $locus=$seq->trunc($start,$end),"\n";# gets sequence of gene from start to end
            location as a SEQ OBJECT
            $locus->display_id($locus_name);
            $prefix=$filename;
            $prefix=~s/.gb$/_gene/;# this will take each file name and assign to new
            variable ($prefix)
            #each file name (variable)is taken and where there is a .gbk at the end
            (/.gbk$/)of the filename it is substituted (s) with nothing (//)
            $locus_seqIO=Bio::SeqIO->new(-file=>">/home/richard/gbs_all/$locus_name.fas",
            -format=>'fasta');
            $locus_seqIO->write_seq($locus);
                }
            }
        }
```

9.8.2 Reciprocal BLAST Script

```
#!/usr/bin/perl
#
#
#
#blast.pl
#Performs recipricol blast

#use lib "/usr/local/share/perl/5.8.8/Bio";
#use lib "/usr/local/share/perl/5.8.8/Bio/Tools/Run";
#use lib "/home/richard/perl_modules/Tools/BPlite/";
use Bio::SeqIO;
use Bio::Tools::Run::StandAloneBlast;
use Bio::Tools::BPlite::HSP;
my $positive=0;
my $negative=0;
chdir "../gbs_all_2/2603VR/";
@fasta_seq=<*.fas>;

FASTA_SEQ_LOOP: foreach my $fasta_seq(@fasta_seq){
    my $seqio=Bio::SeqIO->new ('-format'=>'fasta',-file=>$fasta_seq);
    #seqio:makes new sequence object.  The individual fasta files for subtilis genes are
    used as ref genome sequences.
    my $seq=$seqio->next_seq;
    print $seq->id,"\n";

    #@database=("B_anthracis_ames","B_anthracis_AmesAnt","B_anthracis_Sterne","B_cereus_1098
    7","B_cereus_14579","B_cereus_E33L","B_clausii","B_halodurans","B_licheniformis_AE017333
    ","B_licheniformis_CP000002","B_thuringiensis");
    @database=("A909","NEM316","515","18RS21","CJB111","COH1","H36B");# results_3
    #array of each individual concatanated gene database
    my @subject_hits=();
    foreach $database(@database){
        my $database_path="../gbs_cat_2/$database/";#specifies the path to the database
        my $top_hit=localblast($seq,$database_path);#each gene is then fed into each
        database
        if (!$top_hit){
            print "no hit\n";
            $negative++;
            next FASTA_SEQ_LOOP;
        }
        $hit_length=0;
        $hit_identical_res=0;
        $top_hit_name=$top_hit->name;
        while ($hsp=$top_hit->next_hsp){#get the top HSP from the top hit
            $hsp_length=$hsp->hsp_length;
            $hsp_percent=$hsp->percent_identity;
            $hsp_identical_residues=$hsp_percent/100*$hsp_length;
            $hit_identical_res=$hit_identical_res+$hsp_identical_residues;#adds the no of
            residues for each hsp in one top hit together
            $hit_length=$hit_length+$hsp_length; #adds the lengths of all the HSP for one
            hit together
        }
        $total_percent_homology=$hit_identical_res/$hit_length*100;
        $ref_len=$seq->length;#original ref gene length
        $total_percent_length=$hit_length/$ref_len*100;#work out % length of gene covered

        if ($total_percent_homology>50 && $total_percent_length >70){#if HSP % is more
        than 50 then pass thru loop then pass thru next loop and HSP length is > 70% of
        that of gene then pass thru loop and print
        }
```

```
        else{
            print "percent match error,$top_hit_name\n",
            "$total_percent_homology,$total_percent_length\n";
            $negative++;
            next FASTA_SEQ_LOOP;
        }
        my $subject_gene_file="/home/richard/gbs_all_2/$database/$top_hit_name.fas";
        my $subject_seqio=Bio::SeqIO->new ('-format'=>'fasta',-file=>$subject_gene_file);
        my $subject_seq=$subject_seqio->next_seq;
        my $ref_database="../gbs_cat_2/2603VR";
        $top_hit=localblast($subject_seq,$ref_database);
        $ref_seq_name=$seq->id;
        if ($top_hit_name==$ref_seq_name){
        }
        else {
            print "$top_hit_name,$ref_seq_name\n";#prints this if neg result for reverse
            blast
            $negative++;
            next FASTA_SEQ_LOOP;

        }
        push(@subject_hits,$subject_seq);
    }
    $positive++;
    unshift (@subject_hits,$seq);
    #my $seqout=Bio::SeqIO->new
    ('-format'=>'fasta',-file=>">/home/f0/nadia/results_total_hsp/$ref_seq_name.fas");
    my $seqout=Bio::SeqIO->new ('-format'=>'fasta',-file=>
    ">/home/richard/gbs_core_3/$ref_seq_name.fas");
    foreach $seq_set(@subject_hits){
        $seqout->write_seq($seq_set);
    }
}
print "$positive,$negative\n";
```

```
#localblast script
sub localblast{ #sub:subroutine followed by name u give it eg localblast(this subroutine
name is then used above to run this part of the script)
    my ($ref_seq,$database)=@_;#@_:  takes the input that is going thru the subroutine and
    passes it into the loop
    $database="/home/richard/gbs_cat_2/".$database;#db location
    @parameters= ('program'=> 'blastn', 'database'=> $database,'X'=>150, 'q'=> -1, 'F'=>"F"
    ,_READMETHOD=> 'Blast');
    #@parameters: parameters for the blast
    my $blastobj = Bio::Tools::Run::StandAloneBlast->new (@parameters);
    #defines new blastobj with the parameters defined above
    $blast_report= $blastobj->blastall ($ref_seq);#takes each seq in turn and blasts
    against db
    my $blast_result=$blast_report->next_result;#gets blast result
    $top_hit=$blast_result->next_hit;#takes top hit from blast result
    return $top_hit;
    }
```

9.8.3 Alignment Script

```
#!/usr/bin/perl
#
#
#
#clustal_core.pl
#Performs alignments

use Bio::Tools::Run::Alignment::Clustalw;
chdir"../gbs_shotgun_core_A909ref/";
@core_gene_seq=<*.fas>;
foreach my $core_gene_seq(@core_gene_seq){
    my @prot_obj;
    my %seq_obj;

    my $seqio=Bio::SeqIO->new ('-format'=>'fasta',-file=>$core_gene_seq);#creates seqIO
    object which in this case is the fasta file of each core gene
    while (my $seq=$seqio->next_seq){#while there is a sequence in the file create a new
    seq obj($seq is the seq obj)
        $seq_obj{$seq->id}=$seq;
        $prot_obj=$seq->translate;
        push(@prot_obj,$prot_obj);# push has to be within the while loop so inside the
        curly brackets
    }


    @params=('quiet' =>1);#this will set the parameters for the clustal. if none set then
    will run at default
    $factory=Bio::Tools::Run::Alignment::Clustalw->new(@params);#factory object created
    $prot_obj_ref=\@prot_obj;#array of alignIO objects to be used in the clustal alignment
    $aln=$factory->align($prot_obj_ref);# $aln is a simplealign object

    $CODONSIZE = 3;
    my $dnaalign = new Bio::SimpleAlign;
    my $seqorder = 0;
    my $alnlen = $aln->length;
    foreach my $seq ( $aln->each_seq ) {
        my $newseq;
        $prot_seq_name = $seq->id;
        #print $prot_seq_name,"\n";


        foreach my $pos ( 1..$alnlen ) {
                        my $loc = $seq->location_from_column($pos);
                        my $dna = '';
                        if( !defined $loc || $loc->location_type ne 'EXACT' ) {
                                $dna = '---';
                        } else {
                                # to readjust to codon boundaries
                                # end needs to be +1 so we can just multiply by
                                CODONSIZE
                                # to get this
                                my ($start,$end) = (($loc->start - 1)*$CODONSIZE) +1,(
                                $loc->end)*$CODONSIZE);
                                if( $start <=0 || $end > $seq_obj{$prot_seq_name}->
                                length() ) {
                                        print "start is ", $loc->start, " end is ",
                                         $loc->end, "\n";
                                        warn("codons don't seem to be matching up
                                        for $start,$end");
                                        $dna = '---';
                                    }   else {
```

```
                                                    $dna = $seq_obj{$prot_seq_name}->subseq
                                                    ($start,$end);
                    }
                            }
                            $newseq .= $dna;
                    }
                    $seqorder++;
                    # funky looking math is to readjust to codon boundaries and deal
                    # with fact that sequence start with 1
                    my $newdna = new Bio::LocatableSeq(-display_id  => $seq->id(),
                                 -start => (($seq->start - 1) * $CODONSIZE) + 1,
                                 -end   => ($seq->end * $CODONSIZE),
                                 -strand => $seq->strand,
                                 -seq   => $newseq);
                    $dnaalign->add_seq($newdna);
            }
            my $core_gene_name=$core_gene_seq;
            $core_gene_name=~s/\.fas$//;
            $alignout= Bio::AlignIO->new(-file=>">/home/f0/rloy/shotgun_nexus/$core_gene_name.nex",
            -format =>'nexus');
            $alignout->write_aln($dnaalign);
            print $core_gene_name,"\n";

    }
```

9.8.4 ANI Script

```
#!/usr/bin/perl
#
#
#
#ANI.pl
#Calculates ANI

use Bio::AlignIO;
chdir"../shotgun_distance_corrected";
@nexus=<*.nex>;
%aln_len={};
%identical_bases={};
foreach my $nexus(@nexus){
    print $nexus,"\n";
    $nexus_IO=Bio::AlignIO->new(-file=>"$nexus",-format =>'nexus');
    my $nexus_aln=$nexus_IO->next_aln;
    my $number_of_sequences=$nexus_aln->no_sequences;
    for (my $i=1;$i<$number_of_sequences;$i++){
        for (my $j = $i+1; $j <= $number_of_sequences; $j++) {
                my $seq1= $nexus_aln->get_seq_by_pos($i);

                my $seq2= $nexus_aln->get_seq_by_pos($j);

                my $aln = new Bio::SimpleAlign();

                $aln->add_seq($seq1);

                $aln->add_seq($seq2);

                my $percent_id = $aln -> overall_percentage_identity();

                my $percent_id_output = sprintf "%.2f", $percent_id;

                #print "Percent id between ", $seq1->id , " and " , $seq2->id, " is
                $percent_id_output\n";
                $aln_len=$aln->length;
                $number_identical_bases=($percent_id_output/100)*$aln_len;
                #print "number of identical bases is,$number_identical_bases\n";
                $genome1=$seq1->id;
                $genome2=$seq2->id;
                if ($genome1=~/SAG\d/){
                    $genome1="2603VR";
                }
                elsif ($genome1=~/COH1_\d/){
                    $genome1="COH1";
                }
                elsif ($genome1=~/CJB111_\d/){
                    $genome1="CJB111";
                }
                elsif ($genome1=~/H36B_\d/){
                    $genome1="H36B";
                }
                elsif ($genome1=~/gbs\d/){
                    $genome1="NEM316";
                }
                elsif ($genome1=~/SGB_\d/){
                    $genome1="515";
                }
                else {
                    $genome1="A909";
                }
```

```
                if ($genome2=~/SAG\d/){
                    $genome2="2603VR";
                }
                elsif ($genome2=~/COH1_\d/){
                    $genome2="COH1";
                }
                elsif ($genome2=~/CJB111_\d/){
                    $genome2="CJB111";
                }
                elsif ($genome2=~/H36B_\d/){
                    $genome2="H36B";
                }
                elsif ($genome2=~/gbs\d/){
                    $genome2="NEM316";
                }
                elsif ($genome2=~/SGB_\d/){
                    $genome2="515";
                }
                else {
                    $genome2="A909";
                }
                $aln_len{"$genome1-$genome2"}+=$aln_len;
                $aln_len{"$genome2-$genome1"}+=$aln_len;
                $identical_bases{"$genome1-$genome2"}+=$number_identical_bases;
                $identical_bases{"$genome2-$genome1"}+=$number_identical_bases;

            }

    }

}
    open (OUT,">ANI.txt");
    print OUT "2603VR-COH1\t", 1-($identical_bases{'2603VR-COH1'}/$aln_len{'2603VR-COH1'}),
     "\n";
    print OUT "2603VR-CJB111\t", 1-($identical_bases{'2603VR-CJB111'}/$aln_len{
    '2603VR-CJB111'}), "\n";
    print OUT "2603VR-H36B\t", 1-($identical_bases{'2603VR-H36B'}/$aln_len{'2603VR-H36B'}),
     "\n";
    print OUT "2603VR-NEM316\t", 1-($identical_bases{'2603VR-NEM316'}/$aln_len{
    '2603VR-NEM316'}), "\n";
    print OUT "2603VR-515\t", 1-($identical_bases{'2603VR-515'}/$aln_len{'2603VR-515'}),
    "\n";
    print OUT "2603VR-A909\t", 1-($identical_bases{'2603VR-A909'}/$aln_len{'2603VR-A909'}),
     "\n";
    print OUT "COH1-CJB111\t", 1-($identical_bases{'COH1-CJB111'}/$aln_len{'COH1-CJB111'}),
     "\n";
    print OUT "COH1-H36B\t", 1-($identical_bases{'COH1-H36B'}/$aln_len{'COH1-H36B'}), "\n";
    print OUT "COH1-NEM316\t", 1-($identical_bases{'COH1-NEM316'}/$aln_len{'COH1-NEM316'}),
     "\n";
    print OUT "COH1-H36B\t", 1-($identical_bases{'COH1-H36B'}/$aln_len{'COH1-H36B'}), "\n";
    print OUT "COH1-NEM316\t", 1-($identical_bases{'COH1-NEM316'}/$aln_len{'COH1-NEM316'}),
     "\n";
    print OUT "COH1-515\t", 1-($identical_bases{'COH1-515'}/$aln_len{'COH1-515'}), "\n";
    print OUT "COH1-A909\t", 1-($identical_bases{'COH1-A909'}/$aln_len{'COH1-A909'}), "\n";

    print OUT "CJB111-H36B\t", 1-($identical_bases{'CJB111-H36B'}/$aln_len{'CJB111-H36B'}),
     "\n";
    print OUT "CJB111-NEM316\t", 1-($identical_bases{'CJB111-NEM316'}/$aln_len{
    'CJB111-NEM316'}), "\n";
    print OUT "CJB111-515\t", 1-($identical_bases{'CJB111-515'}/$aln_len{'CJB111-515'}),
    "\n";
```

```
print OUT "CJB111-A909\t", 1-($identical_bases{'CJB111-A909'}/$aln_len{'CJB111-A909'}),
 "\n";
print OUT "H36B-NEM316\t", 1-($identical_bases{'H36B-NEM316'}/$aln_len{'H36B-NEM316'}),
 "\n";
print OUT "H36B-515\t", 1-($identical_bases{'H36B-515'}/$aln_len{'H36B-515'}), "\n";
print OUT "H36B-A909\t", 1-($identical_bases{'H36B-A909'}/$aln_len{'H36B-A909'}), "\n";
print OUT "NEM316-515\t", 1-($identical_bases{'NEM316-515'}/$aln_len{'NEM316-515'}),
"\n";
print OUT "NEM316-A909\t", 1-($identical_bases{'NEM316-A909'}/$aln_len{'NEM316-A909'}),
 "\n";
```

9.8.5 Distance Analysis Script

```
#!/usr/bin/perl
#
#
#
#distance.pl
#Coordinates ModelTest and Paup to perfrom analysis


chdir "/home/f0/rloy/gbs_core_8_trimmed_failed/";
@nexus=<*.nex>;
$dir="/home/f0/rloy/gbs_core_8_trimmed_failed/";
foreach $nexus(@nexus){
    $fileprefix=$nexus;
    $fileprefix=~s/\.nex$/\.scores/;
    if (-e $fileprefix){next;}
    $fileprefix=~s/\.scores/_command\.txt/;
    if (!-e $fileprefix) {
        open (IN,"prefix.txt");
        @prefix=<IN>;
        close IN;
        open(IN,"suffix.txt");
        @suffix=<IN>;
        close IN;
        $nexus=~s/\.nex$/_command\.txt/;
        open (OUT,">$nexus");
        print OUT @prefix;
        $nexus=~s/_command.txt$/\.nex/;
        print OUT "execute $nexus;\n";
        @new_suffix=();
        foreach my $line(@suffix){
            $line=~s/model\.scores/$dir$nexus/;
            $line=~s/\.nex/\.scores/;
            push (@new_suffix,$line);
        }
        print OUT @new_suffix;
        close OUT;
    }
    $nexus=~s/\.nex$/_command\.txt/;
    print "$nexus\n";
    `/usr/local/bin/paup -n < $dir$nexus `;
}

@match_scores=<*.scores>;
foreach $match_scores(@match_scores){
    $model_test=$match_scores;
    $model_test=~s/\.scores/\.out/;
    `modeltest <$dir$match_scores>$model_test`;
    }

@model_test=<*.out>;
foreach $model_test(@model_test){
    open(IN,$model_test);
    until($line=~/AIC/){
        $line=<IN>;
    }
    until($line=~/BEGIN PAUP;/){
        $line=<IN>;
    }
    @params=();
    push(@params,$line);
    $model_test=~s/\.out$/\.nex/;
```

```
    push(@params,"execute $model_test;\n");
    until ($line=~/END/){
        $line=<IN>;
        push(@params,$line);
    }
    $model_test=~s/\.nex$/_params\.txt/;
    pop @params;
    open (OUT,">$model_test");
    print OUT @params;
    print OUT "set criterion=likelihood;\n";
    print OUT "hs;\n";
    $model_test=~s/_params\.txt/\.dist/;
    print OUT "savedist file='/home/f0/rloy/gbs_core_8_trimmed_failed/$model_test';\n";
    print OUT "LOG STOP;\n";
    print OUT "END;\n";
    close OUT;
    $model_test=~s/\.dist/_params\.txt/;
    `/usr/local/bin/paup -n <$dir$model_test> log.file`;
    print "$model_test\n";
}

print "done at last\!\!\!\!\n";
```

9.8.6 Script for Stata Input Formatting

```
#!/usr/bin/perl
#
#
#
#stata.pl
#formats distance data for stata

#chdir 'C:\Modeltest\test\richard_distance_table';
#chdir 'C:\Modeltest\test\concat_align\dist_table';
chdir '/home/f0/rloy/gbs_core_8_trimmed_disttable/';
#open (OUT,">$file_number");
open (IN, "ANI.txt");
@ani=();
while ($line=<IN>){
        chomp $line;
        ($genome_name, $ani)=split(/\t/,$line);
        push(@ani,$ani);
}
@dist_table=<*.dist>;
$counter=0;
open (OT,">file_table.txt");
foreach $dist_table(@dist_table){
    @gene_dist=();
    $counter++;
    print "$counter\n";
    print OT "$counter\t$dist_table\n";
    open (IN, "$dist_table");
    while ($line_2=<IN>){
        chomp $line_2;
        ($gene_name, $gene_dist)=split(/\t/,$line_2);
        push(@gene_dist,$gene_dist);
    }
    #print scalar(@ani);
    open (OUT,">$counter.txt");
    print OUT "gene","\t","genome", "\n";
    for($i=0;$i<scalar(@ani);$i++){
        print OUT "$gene_dist[$i]","\t","$ani[$i]","\n";
    }
}
```

9.8.7 Stata Input Script

```
clear
forval x=1/794{
    clear
    local filename="C:\Modeltest\test\RPL\\`x'.txt"
    insheet using `filename'
    ktau gene genome
    generate file=`x'
    generate tau_a=r(tau_a)
    generate tau_b=r(tau_b)
    keep tau_a tau_b file
    keep if _n==1
    display file
    save "C:\Modeltest\test\RPL\kendall_temp",replace
    if `x'==1{
        outsheet using "C:\Modeltest\test\RPL\kendall_coeff.txt",replace
     }
    else{
        clear
        insheet using "C:\Modeltest\test\RPL\kendall_coeff.txt"
        append using "C:\Modeltest\test\RPL\kendall_temp"
        outsheet using "C:\Modeltest\test\RPL\kendall_coeff.txt",replace
    }
}
```

9.8.8 Script to Calculate Summation of Distance Values

```
#!/usr/bin/perl
#
#
#
#sum_distance.pl
#worksout sum distance for all core genes for ABS calculation.

use warnings;
use strict;

chdir"/home/f0/rloy/gbs_core_8_trimmed_disttable/";

my @distfile = <*.dist>;
my $distfile = (@distfile);


foreach my $distfile(@distfile){

my $output = ">/home/f0/rloy/gbs_core_8_trimmed_disttable/$distfile\_sum";
open (OUTPUT, "$output");
open (DISTANCE, "$distfile");

        while (my $replace = <DISTANCE>){
                $replace =~ s/515-18RS21\s*//g;
                $replace =~ s/515-2603VR\s*//g;
                $replace =~ s/515-A909\s*//g;
                $replace =~ s/515-CJB111\s*//g;
                $replace =~ s/515-COH1\s*//g;
                $replace =~ s/515-H36B\s*//g;
                $replace =~ s/515-NEM316\s*//g;
                $replace =~ s/18RS21-2603VR\s*//g;
                $replace =~ s/18RS21-A909\s*//g;
                $replace =~ s/18RS21-CJB111\s*//g;
                $replace =~ s/18RS21-COH1\s*//g;
                $replace =~ s/18RS21-H36B\s*//g;
                $replace =~ s/18RS21-NEM316\s*//g;
                $replace =~ s/2603VR-A909\s*//g;
                $replace =~ s/2603VR-CJB111\s*//g;
                $replace =~ s/2603VR-COH1\s*//g;
                $replace =~ s/2603VR-H36B\s*//g;
                $replace =~ s/2603VR-NEM316\s*//g;
                $replace =~ s/A909-CJB111\s*//g;
                $replace =~ s/A909-COH1\s*//g;
                $replace =~ s/A909-H36B\s*//g;
                $replace =~ s/A909-NEM316\s*//g;
                $replace =~ s/CJB111-COH1\s*//g;
                $replace =~ s/CJB111-H36B\s*//g;
                $replace =~ s/CJB111-NEM316\s*//g;
                $replace =~ s/COH1-H36B\s*//g;
                $replace =~ s/COH1-NEM316\s*//g;
                $replace =~ s/H36B-NEM316\s*//g;
                $replace =~ s/\n/:/g;

    print OUTPUT ("$replace");
    }
}


my @sum = <*.dist_sum>;
my $sum = @sum;
```

```
my $dist1 = 0;
my $dist2 = 0;
my $dist3 = 0;
my $dist4 = 0;
my $dist5 = 0;
my $dist6 = 0;
my $dist7 = 0;
my $dist8 = 0;
my $dist9 = 0;
my $dist10 = 0;
my $dist11 = 0;
my $dist12 = 0;
my $dist13 = 0;
my $dist14 = 0;
my $dist15 = 0;
my $dist16 = 0;
my $dist17 = 0;
my $dist18 = 0;
my $dist19 = 0;
my $dist20 = 0;
my $dist21 = 0;
my $dist22 = 0;
my $dist23 = 0;
my $dist24 = 0;
my $dist25 = 0;
my $dist26 = 0;
my $dist27 = 0;
my $dist28 = 0;

my @sum_dist = qw//;

foreach $sum(@sum){
open (SUM, "$sum");
while (my $values = <SUM>){
chomp $values;
    ($dist1, $dist2, $dist3, $dist4, $dist5, $dist6, $dist7, $dist8, $dist9, $dist10,
    $dist11, $dist12, $dist13, $dist14, $dist15, $dist16, $dist17, $dist18, $dist19,
    $dist20, $dist21, $dist22, $dist23, $dist24, $dist25, $dist26, $dist27, $dist28) =
    split /:/, $values;
print "$dist1 $dist2 $dist3 $dist4 $dist5 $dist6 $dist7 $dist8 $dist9 $dist10 $dist11
$dist12 $dist13 $dist14 $dist15 $dist16 $dist17 $dist18 $dist19 $dist20 $dist21 $dist22
$dist23 $dist24 $dist25 $dist26 $dist27 $dist28;\n";
my $total=$dist1 + $dist2 + $dist3 + $dist4 + $dist5 + $dist6 + $dist7 + $dist8 + $dist9 +
$dist10 + $dist11 + $dist12 + $dist13 + $dist14 + $dist15 + $dist16 + $dist17 + $dist18 +
$dist19 + $dist20 + $dist21 + $dist22 + $dist23 + $dist24 + $dist25 + $dist26 + $dist27 +
$dist28;
print "$total\n";
push (@sum_dist, $total);

    }
}

foreach my $sum_dist(@sum_dist){

my $sum_output = ">>/home/f0/rloy/gbs_core_8_trimmed_disttable/dist_sum_all.txt";
open (SUM_OUTPUT, "$sum_output");

print SUM_OUTPUT ("$sum_dist\n");
}

foreach $distfile(@distfile){
```

```
my $file_key = ">>/home/f0/rloy/gbs_core_8_trimmed_disttable/file_key.txt";
open (FILE_KEY, "$file_key");

print FILE_KEY ("$distfile\n");
}

`rm *.dist_sum`;

print "Temporary Files Removed\n";
```

9.9 MNR Analysis Scripts

9.9.1 Script to Remove Coding Sequences from the Genome

```
#!/usr/bin/perl
#
#
#
#coding_seq_replace.pl
#Removes Coding Sequences from bacterial genome

#use warnings;
#use strict;

chdir"/home/richard/MNR_final/NEM316/negative_strand/";

my $genome = "/home/richard/MNR_final/NEM316/negative_strand/NEM316_negative_oneline.fasta";
open (GENOME, "$genome");
my $output =
">/home/richard/MNR_final/NEM316/negative_strand/NEM316_noncoding_improved.fasta";
open (OUTPUT, "$output");


my $coding =
"/home/richard/MNR_final/NEM316/negative_strand/NEM316_negative_strand_coding.txt";
open (CODING, "$coding");

my @list = <CODING>;
my $key = 0;
my $value = 0;
my %hash = ();
my $locus = 0;
my $sequence = 0;

foreach my $list(@list){

    ($locus, $sequence) = split /\t/, $list;

    $hash{$locus} = $sequence;

}
while (my $replace = <GENOME>){

    chomp $replace;

    while ((my $key, my $value) = each(%hash)){
        chomp $value;

        my $trim = substr($value, 20);
        my $trim2 = substr($trim, 0, -20);
        $replace =~ s/$trim2/$key/gim;
        $replace =~ s/\n//gim;
        $replace =~ s/\r//gim;

    }
    chomp $replace;
    print OUTPUT ("$replace");
}
```

9.9.2 Script to Parse Non-Coding Regions

```
#!/usr/bin/perl
#
#
#
#get_noncoding.pl
#Quick parser to get each non coding region onto seperate line.

#use warnings;
#use strict;

chdir"/home/richard/MNR_final/NEM316/negative_strand/";

my $genome =
"/home/richard/MNR_final/NEM316/negative_strand/NEM316_noncoding_improved.fasta";
open (GENOME, "$genome");
my $output =
">/home/richard/MNR_final/NEM316/negative_strand/NEM316_negative_noncoding_regions.fasta";
open (OUTPUT, "$output");
my $string = <GENOME>;

while ($string =~ m/(gbs\d\d\d\d[AGCTRYSWKMBDHVN]*gbs\d\d\d\d)/g){

print "$1\n";
print OUTPUT ("$1\n");
}
```

9.9.3 Script to Identify MNR Tracts in Non-Coding DNA

```
#!/usr/bin/perl
#
#
#
#MNR_finder_noncoding.pl
#Identifies Non-Coding regions containing an MNR tract

#use warnings;
#use strict;


chdir"/home/richard/MNR_final/NEM316/negative_strand";

my $noncoding =
"/home/richard/MNR_final/NEM316/negative_strand/NEM316_negative_noncoding_regions.fasta";
open (NONCODING, "$noncoding");
my @array = <NONCODING>;
my $output =
">/home/richard/MNR_final/NEM316/negative_strand/NEM316_negative_noncoding_MNR_new.txt";
open (OUTPUT, $output);

foreach my $sequence(@array){

chomp $sequence;

while($sequence =~
m/(^gbs\d\d\d\d)[AGCTRYSWKMBDHVN]*([TGCN]AAAAAAAAAAAAAAA[TGCN])[AGCTRYSWKMBDHVN]*(gbs\d\d\d\
d)$/gm){print "$1\t15\tA\t$3\t", length ($sequence)-14,"\n"}
while($sequence =~
m/(^gbs\d\d\d\d)[AGCTRYSWKMBDHVN]*([TGCN]AAAAAAAAAAAAAA[TGCN])[AGCTRYSWKMBDHVN]*(gbs\d\d\d\d
)$/gm){print "$1\t14\tA\t$3\t", length ($sequence)-14,"\n"}
while($sequence =~
m/(^gbs\d\d\d\d)[AGCTRYSWKMBDHVN]*([TGCN]AAAAAAAAAAAAA[TGCN])[AGCTRYSWKMBDHVN]*(gbs\d\d\d\d)
$/gm){print "$1\t13\tA\t$3\t", length ($sequence)-14,"\n"}
while($sequence =~
m/(^gbs\d\d\d\d)[AGCTRYSWKMBDHVN]*([TGCN]AAAAAAAAAAAA[TGCN])[AGCTRYSWKMBDHVN]*(gbs\d\d\d\d)$
/gm){print "$1\t12\tA\t$3\t", length ($sequence)-14,"\n"}
while($sequence =~
m/(^gbs\d\d\d\d)[AGCTRYSWKMBDHVN]*([TGCN]AAAAAAAAAA[TGCN])[AGCTRYSWKMBDHVN]*(gbs\d\d\d\d)$/
gm){print "$1\t11\tA\t$3\t", length ($sequence)-14,"\n"}
while($sequence =~
m/(^gbs\d\d\d\d)[AGCTRYSWKMBDHVN]*([TGCN]AAAAAAAAA[TGCN])[AGCTRYSWKMBDHVN]*(gbs\d\d\d\d)$/g
m){print "$1\t10\tA\t$3\t", length ($sequence)-14,"\n"}
while($sequence =~
m/(^gbs\d\d\d\d)[AGCTRYSWKMBDHVN]*([TGCN]AAAAAAAA[TGCN])[AGCTRYSWKMBDHVN]*(gbs\d\d\d\d)$/gm
){print "$1\t9\tA\t$3\t", length ($sequence)-14,"\n"}
while($sequence =~
m/(^gbs\d\d\d\d)[AGCTRYSWKMBDHVN]*([TGCN]AAAAAAA[TGCN])[AGCTRYSWKMBDHVN]*(gbs\d\d\d\d)$/gm
){print "$1\t8\tA\t$3\t", length ($sequence)-14,"\n"}
while($sequence =~
m/(^gbs\d\d\d\d)[AGCTRYSWKMBDHVN]*([TGCN]AAAAAAA[TGCN])[AGCTRYSWKMBDHVN]*(gbs\d\d\d\d)$/gm
){print "$1\t7\tA\t$3\t", length ($sequence)-14,"\n"}
while($sequence =~
m/(^gbs\d\d\d\d)[AGCTRYSWKMBDHVN]*([TGCN]AAAAAA[TGCN])[AGCTRYSWKMBDHVN]*(gbs\d\d\d\d)$/gm){
print "$1\t6\tA\t$3\t", length ($sequence)-14,"\n"}

while($sequence =~
m/(^gbs\d\d\d\d)[AGCTRYSWKMBDHVN]*([AGCN]TTTTTTTTTTTTTTT[AGCN])[AGCTRYSWKMBDHVN]*(gbs\d\d\d\
d)$/gm){print "$1\t15\tT\t$3\t", length ($sequence)-14,"\n"}
while($sequence =~
m/(^gbs\d\d\d\d)[AGCTRYSWKMBDHVN]*([AGCN]TTTTTTTTTTTTTT[AGCN])[AGCTRYSWKMBDHVN]*(gbs\d\d\d\d
```

```
.)$/gm){print "$1\t14\tT\t$3\t", length ($sequence)-14,"\n"}
while($sequence =~
m/(^gbs\d\d\d\d)[AGCTRYSWKMBDHVN]*([AGCN]TTTTTTTTTTTT[AGCN])[AGCTRYSWKMBDHVN]*(gbs\d\d\d\d)
$/gm){print "$1\t13\tT\t$3\t", length ($sequence)-14,"\n"}
while($sequence =~
m/(^gbs\d\d\d\d)[AGCTRYSWKMBDHVN]*([AGCN]TTTTTTTTTTT[AGCN])[AGCTRYSWKMBDHVN]*(gbs\d\d\d\d)$
/gm){print "$1\t12\tT\t$3\t", length ($sequence)-14,"\n"}
while($sequence =~
m/(^gbs\d\d\d\d)[AGCTRYSWKMBDHVN]*([AGCN]TTTTTTTTTT[AGCN])[AGCTRYSWKMBDHVN]*(gbs\d\d\d\d)$/
gm){print "$1\t11\tT\t$3\t", length ($sequence)-14,"\n"}
while($sequence =~
m/(^gbs\d\d\d\d)[AGCTRYSWKMBDHVN]*([AGCN]TTTTTTTTT[AGCN])[AGCTRYSWKMBDHVN]*(gbs\d\d\d\d)$/g
m){print "$1\t10\tT\t$3\t", length ($sequence)-14,"\n"}
while($sequence =~
m/(^gbs\d\d\d\d)[AGCTRYSWKMBDHVN]*([AGCN]TTTTTTTT[AGCN])[AGCTRYSWKMBDHVN]*(gbs\d\d\d\d)$/gm
){print "$1\t9\tT\t$3\t", length ($sequence)-14,"\n"}
while($sequence =~
m/(^gbs\d\d\d\d)[AGCTRYSWKMBDHVN]*([AGCN]TTTTTTT[AGCN])[AGCTRYSWKMBDHVN]*(gbs\d\d\d\d)$/gm
){print "$1\t8\tT\t$3\t", length ($sequence)-14,"\n"}
while($sequence =~
m/(^gbs\d\d\d\d)[AGCTRYSWKMBDHVN]*([AGCN]TTTTTT[AGCN])[AGCTRYSWKMBDHVN]*(gbs\d\d\d\d)$/gm
){print "$1\t7\tT\t$3\t", length ($sequence)-14,"\n"}
while($sequence =~
m/(^gbs\d\d\d\d)[AGCTRYSWKMBDHVN]*([AGCN]TTTTT[AGCN])[AGCTRYSWKMBDHVN]*(gbs\d\d\d\d)$/gm){
print "$1\t6\tT\t$3\t", length ($sequence)-14,"\n"}


while($sequence =~
m/(^gbs\d\d\d\d)[AGCTRYSWKMBDHVN]*([AGTN]CCCCCCCCCCCCCCC[AGTN])[AGCTRYSWKMBDHVN]*(gbs\d\d\d\
d)$/gm){print "$1\t15\tC\t$3\t", length ($sequence)-14,"\n"}
while($sequence =~
m/(^gbs\d\d\d\d)[AGCTRYSWKMBDHVN]*([AGTN]CCCCCCCCCCCCCC[AGTN])[AGCTRYSWKMBDHVN]*(gbs\d\d\d\d
)$/gm){print "$1\t14\tC\t$3\t", length ($sequence)-14,"\n"}
while($sequence =~
m/(^gbs\d\d\d\d)[AGCTRYSWKMBDHVN]*([AGTN]CCCCCCCCCCCCC[AGTN])[AGCTRYSWKMBDHVN]*(gbs\d\d\d\d)
$/gm){print "$1\t13\tC\t$3\t", length ($sequence)-14,"\n"}
while($sequence =~
m/(^gbs\d\d\d\d)[AGCTRYSWKMBDHVN]*([AGTN]CCCCCCCCCCCC[AGTN])[AGCTRYSWKMBDHVN]*(gbs\d\d\d\d)$
/gm){print "$1\t12\tC\t$3\t", length ($sequence)-14,"\n"}
while($sequence =~
m/(^gbs\d\d\d\d)[AGCTRYSWKMBDHVN]*([AGTN]CCCCCCCCCCC[AGTN])[AGCTRYSWKMBDHVN]*(gbs\d\d\d\d)$/
gm){print "$1\t11\tC\t$3\t", length ($sequence)-14,"\n"}
while($sequence =~
m/(^gbs\d\d\d\d)[AGCTRYSWKMBDHVN]*([AGTN]CCCCCCCCCC[AGTN])[AGCTRYSWKMBDHVN]*(gbs\d\d\d\d)$/g
m){print "$1\t10\tC\t$3\t", length ($sequence)-14,"\n"}
while($sequence =~
m/(^gbs\d\d\d\d)[AGCTRYSWKMBDHVN]*([AGTN]CCCCCCCCC[AGTN])[AGCTRYSWKMBDHVN]*(gbs\d\d\d\d)$/gm
){print "$1\t9\tC\t$3\t", length ($sequence)-14,"\n"}
while($sequence =~
m/(^gbs\d\d\d\d)[AGCTRYSWKMBDHVN]*([AGTN]CCCCCCCC[AGTN])[AGCTRYSWKMBDHVN]*(gbs\d\d\d\d)$/gm
){print "$1\t8\tC\t$3\t", length ($sequence)-14,"\n"}
while($sequence =~
m/(^gbs\d\d\d\d)[AGCTRYSWKMBDHVN]*([AGTN]CCCCCCC[AGTN])[AGCTRYSWKMBDHVN]*(gbs\d\d\d\d)$/gm
){print "$1\t7\tC\t$3\t", length ($sequence)-14,"\n"}
while($sequence =~
m/(^gbs\d\d\d\d)[AGCTRYSWKMBDHVN]*([AGTN]CCCCCC[AGTN])[AGCTRYSWKMBDHVN]*(gbs\d\d\d\d)$/gm){
print "$1\t6\tC\t$3\t", length ($sequence)-14,"\n"}


while($sequence =~
m/(^gbs\d\d\d\d)[AGCTRYSWKMBDHVN]*([ACTN]GGGGGGGGGGGGGG[ACTN])[AGCTRYSWKMBDHVN]*(gbs\d\d\d\
```

```
d)$/gm){print "$1\t15\tG\t$3\t", length ($sequence)-14,"\n"}
while($sequence =~
m/(^gbs\d\d\d\d)[AGCTRYSWKMBDHVN]*([ACTN]GGGGGGGGGGGGG[ACTN])[AGCTRYSWKMBDHVN]*(gbs\d\d\d\d
)$/gm){print "$1\t14\tG\t$3\t", length ($sequence)-14,"\n"}
while($sequence =~
m/(^gbs\d\d\d\d)[AGCTRYSWKMBDHVN]*([ACTN]GGGGGGGGGGGG[ACTN])[AGCTRYSWKMBDHVN]*(gbs\d\d\d\d)
$/gm){print "$1\t13\tG\t$3\t", length ($sequence)-14,"\n"}
while($sequence =~
m/(^gbs\d\d\d\d)[AGCTRYSWKMBDHVN]*([ACTN]GGGGGGGGGGG[ACTN])[AGCTRYSWKMBDHVN]*(gbs\d\d\d\d)$
/gm){print "$1\t12\tG\t$3\t", length ($sequence)-14,"\n"}
while($sequence =~
m/(^gbs\d\d\d\d)[AGCTRYSWKMBDHVN]*([ACTN]GGGGGGGGGG[ACTN])[AGCTRYSWKMBDHVN]*(gbs\d\d\d\d)$/
gm){print "$1\t11\tG\t$3\t", length ($sequence)-14,"\n"}
while($sequence =~
m/(^gbs\d\d\d\d)[AGCTRYSWKMBDHVN]*([ACTN]GGGGGGGGG[ACTN])[AGCTRYSWKMBDHVN]*(gbs\d\d\d\d)$/g
m){print "$1\t10\tG\t$3\t", length ($sequence)-14,"\n"}
while($sequence =~
m/(^gbs\d\d\d\d)[AGCTRYSWKMBDHVN]*([ACTN]GGGGGGGGG[ACTN])[AGCTRYSWKMBDHVN]*(gbs\d\d\d\d)$/gm
){print "$1\t9\tG\t$3\t", length ($sequence)-14,"\n"}
while($sequence =~
m/(^gbs\d\d\d\d)[AGCTRYSWKMBDHVN]*([ACTN]GGGGGGGG[ACTN])[AGCTRYSWKMBDHVN]*(gbs\d\d\d\d)$/gm
){print "$1\t8\tG\t$3\t", length ($sequence)-14,"\n"}
while($sequence =~
m/(^gbs\d\d\d\d)[AGCTRYSWKMBDHVN]*([ACTN]GGGGGGG[ACTN])[AGCTRYSWKMBDHVN]*(gbs\d\d\d\d)$/gm
){print "$1\t7\tG\t$3\t", length ($sequence)-14,"\n"}
while($sequence =~
m/(^gbs\d\d\d\d)[AGCTRYSWKMBDHVN]*([AGTN]GGGGGG[ACTN])[AGCTRYSWKMBDHVN]*(gbs\d\d\d\d)$/gm){
print "$1\t6\tG\t$3\t", length ($sequence)-14,"\n"}


##########################################################Bellow=TextfileOutput##Above=ScreenOut####
############################################################


while($sequence =~
m/(^gbs\d\d\d\d)[AGCTRYSWKMBDHVN]*([TGCN]AAAAAAAAAAAAAAA[TGCN])[AGCTRYSWKMBDHVN]*(gbs\d\d\d\
d)$/gm){print OUTPUT ("$1\t15\tA\t$3\t", length ($sequence)-14,"\n")}
while($sequence =~
m/(^gbs\d\d\d\d)[AGCTRYSWKMBDHVN]*([TGCN]AAAAAAAAAAAAAA[TGCN])[AGCTRYSWKMBDHVN]*(gbs\d\d\d\d
)$/gm){print OUTPUT ("$1\t14\tA\t$3\t", length ($sequence)-14,"\n")}
while($sequence =~
m/(^gbs\d\d\d\d)[AGCTRYSWKMBDHVN]*([TGCN]AAAAAAAAAAAAA[TGCN])[AGCTRYSWKMBDHVN]*(gbs\d\d\d\d)
$/gm){print OUTPUT ("$1\t13\tA\t$3\t", length ($sequence)-14,"\n")}
while($sequence =~
m/(^gbs\d\d\d\d)[AGCTRYSWKMBDHVN]*([TGCN]AAAAAAAAAAAA[TGCN])[AGCTRYSWKMBDHVN]*(gbs\d\d\d\d)$
/gm){print OUTPUT ("$1\t12\tA\t$3\t", length ($sequence)-14,"\n")}
while($sequence =~
m/(^gbs\d\d\d\d)[AGCTRYSWKMBDHVN]*([TGCN]AAAAAAAAAA[TGCN])[AGCTRYSWKMBDHVN]*(gbs\d\d\d\d)$/
gm){print OUTPUT ("$1\t11\tA\t$3\t", length ($sequence)-14,"\n")}
while($sequence =~
m/(^gbs\d\d\d\d)[AGCTRYSWKMBDHVN]*([TGCN]AAAAAAAAA[TGCN])[AGCTRYSWKMBDHVN]*(gbs\d\d\d\d)$/g
m){print OUTPUT ("$1\t10\tA\t$3\t", length ($sequence)-14,"\n")}
while($sequence =~
m/(^gbs\d\d\d\d)[AGCTRYSWKMBDHVN]*([TGCN]AAAAAAAA[TGCN])[AGCTRYSWKMBDHVN]*(gbs\d\d\d\d)$/gm
){print OUTPUT ("$1\t9\tA\t$3\t", length ($sequence)-14,"\n")}
while($sequence =~
m/(^gbs\d\d\d\d)[AGCTRYSWKMBDHVN]*([TGCN]AAAAAAA[TGCN])[AGCTRYSWKMBDHVN]*(gbs\d\d\d\d)$/gm
){print OUTPUT ("$1\t8\tA\t$3\t", length ($sequence)-14,"\n")}
while($sequence =~
m/(^gbs\d\d\d\d)[AGCTRYSWKMBDHVN]*([TGCN]AAAAAA[TGCN])[AGCTRYSWKMBDHVN]*(gbs\d\d\d\d)$/gm
```

```
){print OUTPUT ("$1\t7\tA\t$3\t", length ($sequence)-14,"\n")}
while($sequence =~
m/(^gbs\d\d\d\d)[AGCTRYSWKMBDHVN]*([TGCN]AAAAAA[TGCN])[AGCTRYSWKMBDHVN]*(gbs\d\d\d\d)$/gm){
print OUTPUT ("$1\t6\tA\t$3\t", length ($sequence)-14,"\n")}

while($sequence =~
m/(^gbs\d\d\d\d)[AGCTRYSWKMBDHVN]*([AGCN]TTTTTTTTTTTTTT[AGCN])[AGCTRYSWKMBDHVN]*(gbs\d\d\d\
d)$/gm){print OUTPUT ("$1\t15\tT\t$3\t", length ($sequence)-14,"\n")}
while($sequence =~
m/(^gbs\d\d\d\d)[AGCTRYSWKMBDHVN]*([AGCN]TTTTTTTTTTTTT[AGCN])[AGCTRYSWKMBDHVN]*(gbs\d\d\d\d
)$/gm){print OUTPUT ("$1\t14\tT\t$3\t", length ($sequence)-14,"\n")}
while($sequence =~
m/(^gbs\d\d\d\d)[AGCTRYSWKMBDHVN]*([AGCN]TTTTTTTTTTTT[AGCN])[AGCTRYSWKMBDHVN]*(gbs\d\d\d\d)
$/gm){print OUTPUT ("$1\t13\tT\t$3\t", length ($sequence)-14,"\n")}
while($sequence =~
m/(^gbs\d\d\d\d)[AGCTRYSWKMBDHVN]*([AGCN]TTTTTTTTTTT[AGCN])[AGCTRYSWKMBDHVN]*(gbs\d\d\d\d)$
/gm){print OUTPUT ("$1\t12\tT\t$3\t", length ($sequence)-14,"\n")}
while($sequence =~
m/(^gbs\d\d\d\d)[AGCTRYSWKMBDHVN]*([AGCN]TTTTTTTTTT[AGCN])[AGCTRYSWKMBDHVN]*(gbs\d\d\d\d)$/
gm){print OUTPUT ("$1\t11\tT\t$3\t", length ($sequence)-14,"\n")}
while($sequence =~
m/(^gbs\d\d\d\d)[AGCTRYSWKMBDHVN]*([AGCN]TTTTTTTTT[AGCN])[AGCTRYSWKMBDHVN]*(gbs\d\d\d\d)$/g
m){print OUTPUT ("$1\t10\tT\t$3\t", length ($sequence)-14,"\n")}
while($sequence =~
m/(^gbs\d\d\d\d)[AGCTRYSWKMBDHVN]*([AGCN]TTTTTTTT[AGCN])[AGGTRYSWKMBDHVN]*(gbs\d\d\d\d)$/gm
){print OUTPUT ("$1\t9\tT\t$3\t", length ($sequence)-14,"\n")}
while($sequence =~
m/(^gbs\d\d\d\d)[AGCTRYSWKMBDHVN]*([AGCN]TTTTTTT[AGCN])[AGCTRYSWKMBDHVN]*(gbs\d\d\d\d)$/gm
){print OUTPUT ("$1\t8\tT\t$3\t", length ($sequence)-14,"\n")}
while($sequence =~
m/(^gbs\d\d\d\d)[AGCTRYSWKMBDHVN]*([AGCN]TTTTTT[AGCN])[AGCTRYSWKMBDHVN]*(gbs\d\d\d\d)$/gm
){print OUTPUT ("$1\t7\tT\t$3\t", length ($sequence)-14,"\n")}
while($sequence =~
m/(^gbs\d\d\d\d)[AGCTRYSWKMBDHVN]*([AGCN]TTTTTT[AGCN])[AGCTRYSWKMBDHVN]*(gbs\d\d\d\d)$/gm){
print OUTPUT ("$1\t6\tT\t$3\t", length ($sequence)-14,"\n")}


while($sequence =~
m/(^gbs\d\d\d\d)[AGCTRYSWKMBDHVN]*([AGTN]CCCCCCCCCCCCCCC[AGTN])[AGCTRYSWKMBDHVN]*(gbs\d\d\d\
d)$/gm){print OUTPUT ("$1\t15\tC\t$3\t", length ($sequence)-14,"\n")}
while($sequence =~
m/(^gbs\d\d\d\d)[AGCTRYSWKMBDHVN]*([AGTN]CCCCCCCCCCCCCC[AGTN])[AGCTRYSWKMBDHVN]*(gbs\d\d\d\d
)$/gm){print OUTPUT ("$1\t14\tC\t$3\t", length ($sequence)-14,"\n")}
while($sequence =~
m/(^gbs\d\d\d\d)[AGCTRYSWKMBDHVN]*([AGTN]CCCCCCCCCCCCC[AGTN])[AGCTRYSWKMBDHVN]*(gbs\d\d\d\d)
$/gm){print OUTPUT ("$1\t13\tC\t$3\t", length ($sequence)-14,"\n")}
while($sequence =~
m/(^gbs\d\d\d\d)[AGCTRYSWKMBDHVN]*([AGTN]CCCCCCCCCCCC[AGTN])[AGCTRYSWKMBDHVN]*(gbs\d\d\d\d)$
/gm){print OUTPUT ("$1\t12\tC\t$3\t", length ($sequence)-14,"\n")}
while($sequence =~
m/(^gbs\d\d\d\d)[AGCTRYSWKMBDHVN]*([AGTN]CCCCCCCCCC[AGTN])[AGCTRYSWKMBDHVN]*(gbs\d\d\d\d)$/
gm){print OUTPUT ("$1\t11\tC\t$3\t", length ($sequence)-14,"\n")}
while($sequence =~
m/(^gbs\d\d\d\d)[AGCTRYSWKMBDHVN]*([AGTN]CCCCCCCCC[AGTN])[AGCTRYSWKMBDHVN]*(gbs\d\d\d\d)$/g
m){print OUTPUT ("$1\t10\tC\t$3\t", length ($sequence)-14,"\n")}
while($sequence =~
m/(^gbs\d\d\d\d)[AGCTRYSWKMBDHVN]*([AGTN]CCCCCCCC[AGTN])[AGCTRYSWKMBDHVN]*(gbs\d\d\d\d)$/gm
){print OUTPUT ("$1\t9\tC\t$3\t", length ($sequence)-14,"\n")}
while($sequence =~
m/(^gbs\d\d\d\d)[AGCTRYSWKMBDHVN]*([AGTN]CCCCCCC[AGTN])[AGCTRYSWKMBDHVN]*(gbs\d\d\d\d)$/gm
){print OUTPUT ("$1\t8\tC\t$3\t", length ($sequence)-14,"\n")}
```

```
while($sequence =~
m/(^gbs\d\d\d\d)[AGCTRYSWKMBDHVN]*([AGTN]CCCCCCC[AGTN])[AGCTRYSWKMBDHVN]*(gbs\d\d\d\d)$/gm
){print OUTPUT ("$1\t7\tC\t$3\t", length ($sequence)-14,"\n")}
while($sequence =~
m/(^gbs\d\d\d\d)[AGCTRYSWKMBDHVN]*([AGTN]CCCCCC[AGTN])[AGCTRYSWKMBDHVN]*(gbs\d\d\d\d)$/gm){
print OUTPUT ("$1\t6\tC\t$3\t", length ($sequence)-14,"\n")}


while($sequence =~
m/(^gbs\d\d\d\d)[AGCTRYSWKMBDHVN]*([ACTN]GGGGGGGGGGGGGG[ACTN])[AGCTRYSWKMBDHVN]*(gbs\d\d\d\
d)$/gm){print OUTPUT ("$1\t15\tG\t$3\t", length ($sequence)-14,"\n")}
while($sequence =~
m/(^gbs\d\d\d\d)[AGCTRYSWKMBDHVN]*([ACTN]GGGGGGGGGGGGG[ACTN])[AGCTRYSWKMBDHVN]*(gbs\d\d\d\d
)$/gm){print OUTPUT ("$1\t14\tG\t$3\t", length ($sequence)-14,"\n")}
while($sequence =~
m/(^gbs\d\d\d\d)[AGCTRYSWKMBDHVN]*([ACTN]GGGGGGGGGGGG[ACTN])[AGCTRYSWKMBDHVN]*(gbs\d\d\d\d)
$/gm){print OUTPUT ("$1\t13\tG\t$3\t", length ($sequence)-14,"\n")}
while($sequence =~
m/(^gbs\d\d\d\d)[AGCTRYSWKMBDHVN]*([ACTN]GGGGGGGGGGG[ACTN])[AGCTRYSWKMBDHVN]*(gbs\d\d\d\d)$
/gm){print OUTPUT ("$1\t12\tG\t$3\t", length ($sequence)-14,"\n")}
while($sequence =~
m/(^gbs\d\d\d\d)[AGCTRYSWKMBDHVN]*([ACTN]GGGGGGGGGG[ACTN])[AGCTRYSWKMBDHVN]*(gbs\d\d\d\d)$/
gm){print OUTPUT ("$1\t11\tG\t$3\t", length ($sequence)-14,"\n")}
while($sequence =~
m/(^gbs\d\d\d\d)[AGCTRYSWKMBDHVN]*([ACTN]GGGGGGGGG[ACTN])[AGCTRYSWKMBDHVN]*(gbs\d\d\d\d)$/g
m){print OUTPUT ("$1\t10\tG\t$3\t", length ($sequence)-14,"\n")}
while($sequence =~
m/(^gbs\d\d\d\d)[AGCTRYSWKMBDHVN]*([ACTN]GGGGGGGG[ACTN])[AGCTRYSWKMBDHVN]*(gbs\d\d\d\d)$/gm
){print OUTPUT ("$1\t9\tG\t$3\t", length ($sequence)-14,"\n")}
while($sequence =~
m/(^gbs\d\d\d\d)[AGCTRYSWKMBDHVN]*([ACTN]GGGGGGG[ACTN])[AGCTRYSWKMBDHVN]*(gbs\d\d\d\d)$/gm
){print OUTPUT ("$1\t8\tG\t$3\t", length ($sequence)-14,"\n")}
while($sequence =~
m/(^gbs\d\d\d\d)[AGCTRYSWKMBDHVN]*([ACTN]GGGGGG[ACTN])[AGCTRYSWKMBDHVN]*(gbs\d\d\d\d)$/gm
){print OUTPUT ("$1\t7\tG\t$3\t", length ($sequence)-14,"\n")}
while($sequence =~
m/(^gbs\d\d\d\d)[AGCTRYSWKMBDHVN]*([AGTN]GGGGGG[ACTN])[AGCTRYSWKMBDHVN]*(gbs\d\d\d\d)$/gm){
print OUTPUT ("$1\t6\tG\t$3\t", length ($sequence)-14,"\n")}

}
```

9.9.4 Script to Identify MNR Tracts in Coding DNA

```
#!/usr/bin/perl
#
#
#
#MNR_finder_coding.pl
#Identifies MNR tracts in Coding regions

#use warnings;
#use strict;

chdir"/home/richard/gbs_all/2603VR/";

my @coding = <*.fas>;
#my $coding = (@coding);


foreach my $coding(@coding){

my $output1 = ">>/home/richard/MNR_final/coding/2603VR/coding_repeats/2603VR_all";
open (OUTPUT1, "$output1");
open (CODING, "$coding");

        while (my $replace = <CODING>){

                $replace =~ s/>//gm;
                $replace =~ s/ Streptococcus agalactiae 2603V\/R, complete
                genome.\s*\n/\t/gm;
                $replace =~ s/ /\t/gm;
                $replace =~ s/\n//gm;
                $replace =~ s/\r//gm;
                $replace =~ s/SAG/\nSAG/gm;

    print OUTPUT1 ("$replace");
    }
}

##################################

chdir"/home/richard/MNR_final/coding/2603VR/coding_repeats/";

my $list = "2603VR_all";
open (HASH_FILE, "$list");

my @hash = <HASH_FILE>;
my $key = 0;
my $value = 0;


foreach my $hash(@hash){

my @split = split(/\t/, $hash);

##print "@split[0],@split[1]\n";

my %hash = ();
my $locus = @split[0];
my $sequence = @split[1];

$hash{$locus} = $sequence;

while ((my $key, my $value) = each(%hash)){
```

```
#print "$key\t$value";
my $output = ">/home/richard/MNR_final/coding/2603VR/coding_repeats_length/$locus";
open (OUTPUT, "$output");

    print OUTPUT ("Locus Tag\tNucleotide\tRepeat Length\tPosition\tGene Length\n");
    while($sequence =~ m/[AGCN\s]TTTTTTTTTTTTTTT[AGCN\s]/gm) {print OUTPUT (
    "$locus\tT\t15\t", pos ($sequence)-15,"\t", length ($sequence),"\n")}
    while($sequence =~ m/[AGCN\s]TTTTTTTTTTTTTT[AGCN\s]/gm) {print OUTPUT (
    "$locus\tT\t14\t", pos ($sequence)-14,"\t", length ($sequence),"\n")}
    while($sequence =~ m/[AGCN\s]TTTTTTTTTTTTT[AGCN\s]/gm) {print OUTPUT ("$locus\tT\t13\t"
    , pos ($sequence)-13,"\t", length ($sequence),"\n")}
    while($sequence =~ m/[AGCN\s]TTTTTTTTTTTT[AGCN\s]/gm) {print OUTPUT ("$locus\tT\t12\t",
     pos ($sequence)-12,"\t", length ($sequence),"\n")}
    while($sequence =~ m/[AGCN\s]TTTTTTTTTT[AGCN\s]/gm) {print OUTPUT ("$locus\tT\t11\t",
    pos ($sequence)-11,"\t", length ($sequence),"\n")}
    while($sequence =~ m/[AGCN\s]TTTTTTTTT[AGCN\s]/gm) {print OUTPUT ("$locus\tT\t10\t",
    pos ($sequence)-10,"\t", length ($sequence),"\n")}
    while($sequence =~ m/[AGCN\s]TTTTTTTT[AGCN\s]/gm) {print OUTPUT ("$locus\tT\t9\t", pos
     ($sequence)-9,"\t", length ($sequence),"\n")}
    while($sequence =~ m/[AGCN\s]TTTTTTT[AGCN\s]/gm) {print OUTPUT ("$locus\tT\t8\t", pos
    ($sequence)-8,"\t", length ($sequence),"\n")}
    while($sequence =~ m/[AGCN\s]TTTTTT[AGCN\s]/gm) {print OUTPUT ("$locus\tT\t7\t", pos (
    $sequence)-7,"\t", length ($sequence),"\n")}
    while($sequence =~ m/[AGCN\s]TTTTT[AGCN\s]/gm) {print OUTPUT ("$locus\tT\t6\t", pos (
    $sequence)-6,"\t", length ($sequence),"\n")}
    #while($sequence =~ m/[AGCN\s]TTTT[AGCN\s]/gm) {print OUTPUT ("$locus\tT\t5\t", pos
    ($sequence)-5,"\t", length ($sequence),"\n")}
    #while($sequence =~ m/[AGCN\s]TTT[AGCN\s]/gm) {print OUTPUT ("$locus\tT\t4\t", pos
    ($sequence)-4,"\t", length ($sequence),"\n")}
    #while($sequence =~ m/[AGCN\s]TT[AGCN\s]/gm) {print OUTPUT ("$locus\tT\t3\t", pos
    ($sequence)-3,"\t", length ($sequence),"\n")}

    while($sequence =~ m/[TGCN\s]AAAAAAAAAAAAAAA[TGCN\s]/gm) {print OUTPUT (
    "$locus\tA\t15\t", pos ($sequence)-15,"\t", length ($sequence),"\n")}
    while($sequence =~ m/[TGCN\s]AAAAAAAAAAAAAA[TGCN\s]/gm) {print OUTPUT (
    "$locus\tA\t14\t", pos ($sequence)-14,"\t", length ($sequence),"\n")}
    while($sequence =~ m/[TGCN\s]AAAAAAAAAAAAA[TGCN\s]/gm) {print OUTPUT ("$locus\tA\t13\t"
    , pos ($sequence)-13,"\t", length ($sequence),"\n")}
    while($sequence =~ m/[TGCN\s]AAAAAAAAAAAA[TGCN\s]/gm) {print OUTPUT ("$locus\tA\t12\t",
     pos ($sequence)-12,"\t", length ($sequence),"\n")}
    while($sequence =~ m/[TGCN\s]AAAAAAAAAA[TGCN\s]/gm) {print OUTPUT ("$locus\tA\t11\t",
    pos ($sequence)-11,"\t", length ($sequence),"\n")}
    while($sequence =~ m/[TGCN\s]AAAAAAAAA[TGCN\s]/gm) {print OUTPUT ("$locus\tA\t10\t",
    pos ($sequence)-10,"\t", length ($sequence),"\n")}
    while($sequence =~ m/[TGCN\s]AAAAAAAA[TGCN\s]/gm) {print OUTPUT ("$locus\tA\t9\t", pos
     ($sequence)-9,"\t", length ($sequence),"\n")}
    while($sequence =~ m/[TGCN\s]AAAAAAA[TGCN\s]/gm) {print OUTPUT ("$locus\tA\t8\t", pos
    ($sequence)-8,"\t", length ($sequence),"\n")}
    while($sequence =~ m/[TGCN\s]AAAAAA[TGCN\s]/gm) {print OUTPUT ("$locus\tA\t7\t", pos (
    $sequence)-7,"\t", length ($sequence),"\n")}
    while($sequence =~ m/[TGCN\s]AAAAA[TGCN\s]/gm) {print OUTPUT ("$locus\tA\t6\t", pos (
    $sequence)-6,"\t", length ($sequence),"\n")}
    #while($sequence =~ m/[TGCN\s]AAAA[TGCN\s]/gm) {print OUTPUT ("$locus\tA\t5\t", pos
    ($sequence)-5,"\t", length ($sequence),"\n")}
    #while($sequence =~ m/[TGCN\s]AAA[TGCN\s]/gm) {print OUTPUT ("$locus\tA\t4\t", pos
    ($sequence)-4,"\t", length ($sequence),"\n")}
    #while($sequence =~ m/[TGCN\s]AA[TGCN\s]/gm) {print OUTPUT ("$locus\tA\t3\t", pos
    ($sequence)-3,"\t", length ($sequence),"\n")}

    while($sequence =~ m/[TAGN\s]CCCCCCCCCCCCCCC[TAGN\s]/gm) {print OUTPUT (
```

```
"$locus\tC\t15\t", pos ($sequence)-15,"\t", length ($sequence),"\n")}
while($sequence =~ m/[TAGN\s]CCCCCCCCCCCCC[TAGN\s]/gm) {print OUTPUT (
"$locus\tC\t14\t", pos ($sequence)-14,"\t", length ($sequence),"\n")}
while($sequence =~ m/[TAGN\s]CCCCCCCCCCCC[TAGN\s]/gm) {print OUTPUT ("$locus\tC\t13\t"
, pos ($sequence)-13,"\t", length ($sequence),"\n")}
while($sequence =~ m/[TAGN\s]CCCCCCCCCCC[TAGN\s]/gm) {print OUTPUT ("$locus\tC\t12\t",
 pos ($sequence)-12,"\t", length ($sequence),"\n")}
while($sequence =~ m/[TAGN\s]CCCCCCCCCC[TAGN\s]/gm) {print OUTPUT ("$locus\tC\t11\t",
pos ($sequence)-11,"\t", length ($sequence),"\n")}
while($sequence =~ m/[TAGN\s]CCCCCCCCC[TAGN\s]/gm) {print OUTPUT ("$locus\tC\t10\t",
pos ($sequence)-10,"\t", length ($sequence),"\n")}
while($sequence =~ m/[TAGN\s]CCCCCCCC[TAGN\s]/gm) {print OUTPUT ("$locus\tC\t9\t", pos
 ($sequence)-9,"\t", length ($sequence),"\n")}
while($sequence =~ m/[TAGN\s]CCCCCCC[TAGN\s]/gm) {print OUTPUT ("$locus\tC\t8\t", pos
($sequence)-8,"\t", length ($sequence),"\n")}
while($sequence =~ m/[TAGN\s]CCCCCC[TAGN\s]/gm) {print OUTPUT ("$locus\tC\t7\t", pos (
$sequence)-7,"\t", length ($sequence),"\n")}
while($sequence =~ m/[TAGN\s]CCCCC[TAGN\s]/gm) {print OUTPUT ("$locus\tC\t6\t", pos (
$sequence)-6,"\t", length ($sequence),"\n")}
#while($sequence =~ m/[TAGN\s]CCCC[TAGN\s]/gm) {print OUTPUT ("$locus\tC\t5\t", pos
($sequence)-5,"\t", length ($sequence),"\n")}
#while($sequence =~ m/[TAGN\s]CCC[TAGN\s]/gm) {print OUTPUT ("$locus\tC\t4\t", pos
($sequence)-4,"\t", length ($sequence),"\n")}
#while($sequence =~ m/[TAGN\s]CC[TAGN\s]/gm) {print OUTPUT ("$locus\tC\t3\t", pos
($sequence)-3,"\t", length ($sequence),"\n")}

while($sequence =~ m/[TACN\s]GGGGGGGGGGGGGG[TACN\s]/gm) {print OUTPUT (
"$locus\tG\t15\t", pos ($sequence)-15,"\t", length ($sequence),"\n")}
while($sequence =~ m/[TACN\s]GGGGGGGGGGGGG[TACN\s]/gm) {print OUTPUT (
"$locus\tG\t14\t", pos ($sequence)-14,"\t", length ($sequence),"\n")}
while($sequence =~ m/[TACN\s]GGGGGGGGGGGG[TACN\s]/gm) {print OUTPUT ("$locus\tG\t13\t"
, pos ($sequence)-13,"\t", length ($sequence),"\n")}
while($sequence =~ m/[TACN\s]GGGGGGGGGGG[TACN\s]/gm) {print OUTPUT ("$locus\tG\t12\t",
 pos ($sequence)-12,"\t", length ($sequence),"\n")}
while($sequence =~ m/[TACN\s]GGGGGGGGGG[TACN\s]/gm) {print OUTPUT ("$locus\tG\t11\t",
pos ($sequence)-11,"\t", length ($sequence),"\n")}
while($sequence =~ m/[TACN\s]GGGGGGGGG[TACN\s]/gm) {print OUTPUT ("$locus\tG\t10\t",
pos ($sequence)-10,"\t", length ($sequence),"\n")}
while($sequence =~ m/[TACN\s]GGGGGGGG[TACN\s]/gm) {print OUTPUT ("$locus\tG\t9\t", pos
 ($sequence)-9,"\t", length ($sequence),"\n")}
while($sequence =~ m/[TACN\s]GGGGGGG[TACN\s]/gm) {print OUTPUT ("$locus\tG\t8\t", pos
($sequence)-8,"\t", length ($sequence),"\n")}
while($sequence =~ m/[TACN\s]GGGGGG[TACN\s]/gm) {print OUTPUT ("$locus\tG\t7\t", pos (
$sequence)-7,"\t", length ($sequence),"\n")}
while($sequence =~ m/[TACN\s]GGGGG[TACN\s]/gm) {print OUTPUT ("$locus\tG\t6\t", pos (
$sequence)-6,"\t", length ($sequence),"\n")}
#while($sequence =~ m/[TACN\s]GGGG[TACN\s]/gm) {print OUTPUT ("$locus\tG\t5\t", pos
($sequence)-5,"\t", length ($sequence),"\n")}
#while($sequence =~ m/[TACN\s]GGG[TACN\s]/gm) {print OUTPUT ("$locus\tG\t4\t", pos
($sequence)-4,"\t", length ($sequence),"\n")}
#while($sequence =~ m/[TACN\s]GG[TACN\s]/gm) {print OUTPUT ("$locus\tG\t3\t", pos
($sequence)-3,"\t", length ($sequence),"\n")}

}
}
```

9.10 GBS Core Genomes

9.10.1 Three Genome Dataset Core Genome

Locus Tag	Gi	Locus Tag	Gi	Locus Tag	Gi	Locus Tag	Gi
SAG0001	22536186	SAG0071	22536256	SAG0143	22536328	SAG0215	22536399
SAG0002	22536187	SAG0072	22536257	SAG0144	22536329	SAG0217	22536401
SAG0003	22536188	SAG0073	22536258	SAG0145	22536330	SAG0220	22536404
SAG0004	22536189	SAG0074	22536259	SAG0146	22536331	SAG0222	22536406
SAG0006	22536191	SAG0075	22536260	SAG0147	22536332	SAG0223	22536407
SAG0007	22536192	SAG0076	22536261	SAG0148	22536333	SAG0224	22536408
SAG0008	22536193	SAG0077	22536262	SAG0149	22536334	SAG0225	22536409
SAG0010	22536195	SAG0078	22536263	SAG0150	22536335	SAG0226	22536410
SAG0011	22536196	SAG0079	22536264	SAG0151	22536336	SAG0227	22536411
SAG0012	22536197	SAG0080	22536265	SAG0152	22536337	SAG0228	22536412
SAG0013	22536198	SAG0082	22536267	SAG0153	22536338	SAG0229	22536413
SAG0014	22536199	SAG0083	161485618	SAG0154	22536339	SAG0230	22536414
SAG0015	22536200	SAG0084	22536269	SAG0155	22536340	SAG0231	22536415
SAG0016	22536201	SAG0085	22536270	SAG0156	22536341	SAG0232	22536416
SAG0017	22536202	SAG0086	22536271	SAG0158	22536342	SAG0234	22536418
SAG0018	22536203	SAG0089	22536274	SAG0159	22536343	SAG0235	22536419
SAG0019	22536204	SAG0090	22536275	SAG0160	22536344	SAG0239	22536423
SAG0020	22536205	SAG0091	22536276	SAG0161	22536345	SAG0240	22536424
SAG0021	22536206	SAG0092	22536277	SAG0162	22536346	SAG0241	22536425
SAG0022	22536207	SAG0093	22536278	SAG0163	22536347	SAG0242	22536426
SAG0023	22536208	SAG0094	22536279	SAG0164	22536348	SAG0243	22536427
SAG0024	22536209	SAG0095	22536280	SAG0165	22536349	SAG0244	22536428
SAG0025	22536210	SAG0096	22536281	SAG0166	22536350	SAG0252	22536436
SAG0026	22536211	SAG0097	22536282	SAG0167	22536351	SAG0253	22536437
SAG0027	22536212	SAG0098	22536283	SAG0168	22536352	SAG0254	22536438
SAG0028	22536213	SAG0099	22536284	SAG0169	22536353	SAG0255	22536439
SAG0029	22536214	SAG0100	22536285	SAG0171	22536355	SAG0256	22536440
SAG0030	22536215	SAG0101	22536286	SAG0172	22536356	SAG0257	22536441
SAG0031	22536216	SAG0102	22536287	SAG0173	22536357	SAG0258	22536442
SAG0032	22536217	SAG0103	22536288	SAG0174	22536358	SAG0259	22536443
SAG0033	22536218	SAG0104	22536289	SAG0175	22536359	SAG0260	22536444
SAG0034	22536219	SAG0105	22536290	SAG0176	22536360	SAG0263	22536447
SAG0035	22536220	SAG0106	22536291	SAG0177	22536361	SAG0264	22536448
SAG0036	22536221	SAG0107	22536292	SAG0178	22536362	SAG0265	22536449
SAG0037	22536222	SAG0108	22536293	SAG0179	22536363	SAG0266	22536450
SAG0038	22536223	SAG0109	22536294	SAG0180	22536364	SAG0267	22536451
SAG0039	22536224	SAG0110	22536295	SAG0181	22536365	SAG0268	22536452
SAG0040	22536225	SAG0111	22536296	SAG0182	22536366	SAG0269	22536453
SAG0041	22536226	SAG0112	22536297	SAG0183	22536367	SAG0270	22536454
SAG0042	22536227	SAG0113	22536298	SAG0184	22536368	SAG0271	22536455
SAG0043	22536228	SAG0114	22536299	SAG0185	22536369	SAG0272	22536456
SAG0044	22536229	SAG0115	22536300	SAG0187	22536371	SAG0273	22536457
SAG0045	22536230	SAG0116	22536301	SAG0188	22536372	SAG0274	22536458
SAG0047	22536232	SAG0117	22536302	SAG0189	22536373	SAG0275	22536459
SAG0049	22536234	SAG0118	22536303	SAG0190	22536374	SAG0276	22536460
SAG0050	22536235	SAG0119	22536304	SAG0191	22536375	SAG0277	22536461
SAG0051	22536236	SAG0121	22536306	SAG0192	22536376	SAG0278	22536462
SAG0052	22536237	SAG0122	22536307	SAG0193	22536377	SAG0279	22536463
SAG0053	22536238	SAG0123	22536308	SAG0194	22536378	SAG0280	22536464
SAG0054	22536239	SAG0124	22536309	SAG0196	22536380	SAG0281	22536465
SAG0055	22536240	SAG0125	22536310	SAG0197	22536381	SAG0282	22536466
SAG0056	22536241	SAG0126	22536311	SAG0198	22536382	SAG0283	22536467
SAG0057	22536242	SAG0127	22536312	SAG0199	22536383	SAG0284	22536468
SAG0058	22536243	SAG0128	22536313	SAG0200	22536384	SAG0285	161485617
SAG0059	22536244	SAG0129	22536314	SAG0201	22536385	SAG0286	22536470
SAG0060	22536245	SAG0130	22536315	SAG0202	22536386	SAG0287	22536471
SAG0061	22536246	SAG0131	22536316	SAG0203	22536387	SAG0288	22536472
SAG0062	22536247	SAG0132	22536317	SAG0204	22536388	SAG0289	22536473
SAG0063	22536248	SAG0135	22536320	SAG0205	22536389	SAG0290	22536474
SAG0064	22536249	SAG0136	22536321	SAG0206	22536390	SAG0291	22536475
SAG0065	22536250	SAG0137	22536322	SAG0207	22536391	SAG0292	22536476
SAG0066	22536251	SAG0138	22536323	SAG0208	22536392	SAG0293	22536477
SAG0067	22536252	SAG0139	22536324	SAG0209	22536393	SAG0294	22536478
SAG0068	22536253	SAG0140	22536325	SAG0210	22536394	SAG0295	22536479
SAG0069	22536254	SAG0141	22536326	SAG0211	22536395	SAG0296	22536480
SAG0070	22536255	SAG0142	22536327	SAG0214	22536398	SAG0297	22536481

Locus Tag	GI
SAG0298	22536482
SAG0299	22536483
SAG0300	22536484
SAG0302	22536486
SAG0303	22536487
SAG0304	22536488
SAG0305	22536489
SAG0306	22536490
SAG0308	22536492
SAG0309	22536493
SAG0310	22536494
SAG0312	22536495
SAG0313	22536496
SAG0314	22536497
SAG0315	22536498
SAG0316	22536499
SAG0317	22536500
SAG0318	22536501
SAG0319	22536502
SAG0320	22536503
SAG0321	22536504
SAG0322	22536505
SAG0323	22536506
SAG0324	22536507
SAG0325	22536508
SAG0326	22536509
SAG0327	22536510
SAG0328	22536511
SAG0329	22536512
SAG0330	22536513
SAG0331	22536514
SAG0332	22536515
SAG0333	22536516
SAG0334	22536517
SAG0335	22536518
SAG0336	22536519
SAG0337	22536520
SAG0338	22536521
SAG0339	22536522
SAG0340	22536523
SAG0342	22536525
SAG0343	22536526
SAG0344	22536527
SAG0345	22536528
SAG0346	22536529
SAG0347	22536530
SAG0348	22536531
SAG0349	22536532
SAG0350	22536533
SAG0351	22536534
SAG0352	22536535
SAG0353	22536536
SAG0354	22536537
SAG0355	22536538
SAG0356	22536539
SAG0357	22536540
SAG0358	22536541
SAG0359	22536542
SAG0360	22536543
SAG0361	22536544
SAG0362	22536545
SAG0363	22536546
SAG0364	22536547
SAG0365	22536548
SAG0366	22536549
SAG0367	22536550

Locus Tag	GI
SAG0368	22536551
SAG0369	22536552
SAG0370	22536553
SAG0371	22536554
SAG0373	22536556
SAG0374	22536557
SAG0375	22536558
SAG0376	22536559
SAG0377	22536560
SAG0378	161485616
SAG0379	22536562
SAG0380	22536563
SAG0381	22536564
SAG0382	22536565
SAG0383	22536566
SAG0384	22536567
SAG0385	22536568
SAG0386	22536569
SAG0387	22536570
SAG0388	22536571
SAG0389	22536572
SAG0390	22536573
SAG0391	22536574
SAG0392	22536575
SAG0393	22536576
SAG0394	22536577
SAG0395	22536578
SAG0396	22536579
SAG0397	22536580
SAG0398	22536581
SAG0399	22536582
SAG0400	22536583
SAG0402	22536585
SAG0403	22536586
SAG0404	22536587
SAG0405	22536588
SAG0406	22536589
SAG0407	22536590
SAG0408	22536591
SAG0409	22536592
SAG0410	22536593
SAG0411	22536594
SAG0412	22536595
SAG0413	22536596
SAG0414	22536597
SAG0415	22536598
SAG0417	22536600
SAG0418	22536601
SAG0419	22536602
SAG0420	22536603
SAG0421	22536604
SAG0422	22536605
SAG0423	22536606
SAG0425	22536608
SAG0426	22536609
SAG0430	22536613
SAG0431	22536614
SAG0432	22536615
SAG0440	22536620
SAG0441	22536621
SAG0442	22536622
SAG0443	22536623
SAG0445	22536625
SAG0446	22536626
SAG0447	22536627
SAG0449	22536629

Locus Tag	GI
SAG0450	22536630
SAG0451	22536631
SAG0452	22536632
SAG0454	22536634
SAG0455	22536635
SAG0457	22536636
SAG0458	22536637
SAG0459	22536638
SAG0460	22536639
SAG0461	22536640
SAG0462	22536641
SAG0463	22536642
SAG0464	22536643
SAG0465	22536644
SAG0467	22536646
SAG0468	22536647
SAG0469	22536648
SAG0470	22536649
SAG0471	22536650
SAG0472	22536651
SAG0473	22536652
SAG0474	22536653
SAG0475	22536654
SAG0476	22536655
SAG0477	22536656
SAG0478	22536657
SAG0479	22536658
SAG0480	22536659
SAG0481	22536660
SAG0482	22536661
SAG0483	22536662
SAG0484	22536663
SAG0485	22536664
SAG0486	22536665
SAG0487	22536666
SAG0488	22536667
SAG0490	22536669
SAG0491	22536670
SAG0492	22536671
SAG0493	22536672
SAG0494	22536673
SAG0495	22536674
SAG0496	22536675
SAG0497	22536676
SAG0498	22536677
SAG0499	22536678
SAG0500	22536679
SAG0501	22536680
SAG0502	22536681
SAG0503	22536682
SAG0504	22536683
SAG0505	22536684
SAG0506	22536685
SAG0507	22536686
SAG0508	22536687
SAG0509	22536688
SAG0510	22536689
SAG0511	22536690
SAG0512	22536691
SAG0513	22536692
SAG0514	22536693
SAG0516	22536695
SAG0517	22536696
SAG0519	22536697
SAG0520	22536698
SAG0521	22536699

Locus Tag	GI
SAG0522	22536700
SAG0523	22536701
SAG0524	22536702
SAG0525	22536703
SAG0526	22536704
SAG0527	22536705
SAG0528	22536706
SAG0530	22536708
SAG0531	22536709
SAG0532	22536710
SAG0533	22536711
SAG0534	22536712
SAG0535	22536713
SAG0536	22536714
SAG0537	22536715
SAG0538	22536716
SAG0539	22536717
SAG0540	22536718
SAG0541	22536719
SAG0544	22536722
SAG0566	22536744
SAG0606	22536782
SAG0610	22536785
SAG0612	22536786
SAG0613	22536787
SAG0614	22536788
SAG0615	22536789
SAG0616	22536790
SAG0617	22536791
SAG0619	22536792
SAG0621	22536794
SAG0622	22536795
SAG0623	22536796
SAG0624	22536797
SAG0625	22536798
SAG0626	22536799
SAG0627	22536800
SAG0628	22536801
SAG0629	22536802
SAG0630	22536803
SAG0631	22536804
SAG0632	22536805
SAG0633	22536806
SAG0634	22536807
SAG0635	22536808
SAG0636	22536809
SAG0640	22536811
SAG0645	22536814
SAG0646	22536815
SAG0647	22536816
SAG0648	22536817
SAG0649	22536818
SAG0651	22536820
SAG0657	22536824
SAG0658	22536825
SAG0659	22536826
SAG0660	22536827
SAG0661	22536828
SAG0662	22536829
SAG0663	22536830
SAG0664	22536831
SAG0665	22536832
SAG0666	22536833
SAG0667	22536834
SAG0668	22536835
SAG0669	22536836

Locus Tag	Gi
SAG0670	22536837
SAG0671	22536838
SAG0672	22536839
SAG0673	22536840
SAG0684	22536849
SAG0685	22536850
SAG0686	22536851
SAG0687	22536852
SAG0688	22536853
SAG0689	22536854
SAG0690	22536855
SAG0691	22536856
SAG0692	22536857
SAG0695	22536859
SAG0696	22536860
SAG0697	22536861
SAG0698	22536862
SAG0699	22536863
SAG0700	22536864
SAG0701	22536865
SAG0702	22536866
SAG0703	22536867
SAG0704	22536868
SAG0705	22536869
SAG0706	22536870
SAG0707	22536871
SAG0708	22536872
SAG0709	22536873
SAG0710	22536874
SAG0711	22536875
SAG0712	22536876
SAG0713	22536877
SAG0714	22536878
SAG0715	22536879
SAG0716	22536880
SAG0717	22536881
SAG0718	22536882
SAG0719	22536883
SAG0720	22536884
SAG0721	22536885
SAG0722	22536886
SAG0723	22536887
SAG0724	22536888
SAG0725	22536889
SAG0726	22536890
SAG0727	22536891
SAG0728	22536892
SAG0729	22536893
SAG0730	22536894
SAG0731	22536895
SAG0732	22536896
SAG0733	22536897
SAG0734	22536898
SAG0736	22536900
SAG0737	22536901
SAG0738	22536902
SAG0739	22536903
SAG0740	22536904
SAG0741	22536905
SAG0742	22536906
SAG0743	22536907
SAG0744	22536908
SAG0745	22536909
SAG0746	22536910
SAG0747	161485615
SAG0748	22536912

Locus Tag	Gi
SAG0749	22536913
SAG0750	22536914
SAG0751	22536915
SAG0752	22536916
SAG0753	22536917
SAG0754	22536918
SAG0755	22536919
SAG0756	22536920
SAG0757	22536921
SAG0758	22536922
SAG0759	22536923
SAG0761	22536925
SAG0762	22536926
SAG0763	22536927
SAG0764	22536928
SAG0765	22536929
SAG0766	22536930
SAG0767	22536931
SAG0768	22536932
SAG0769	22536933
SAG0770	22536934
SAG0771	22536935
SAG0772	22536936
SAG0773	22536937
SAG0774	22536938
SAG0775	22536939
SAG0776	22536940
SAG0777	22536941
SAG0778	22536942
SAG0779	22536943
SAG0780	22536944
SAG0781	22536945
SAG0782	22536946
SAG0783	22536947
SAG0784	22536948
SAG0785	22536949
SAG0786	22536950
SAG0787	22536951
SAG0788	22536952
SAG0789	22536953
SAG0790	22536954
SAG0791	22536955
SAG0792	22536956
SAG0793	22536957
SAG0794	22536958
SAG0795	22536959
SAG0796	22536960
SAG0797	22536961
SAG0798	22536962
SAG0799	22536963
SAG0800	22536964
SAG0801	22536965
SAG0803	22536967
SAG0804	22536968
SAG0805	22536969
SAG0806	22536970
SAG0807	22536971
SAG0808	22536972
SAG0809	22536973
SAG0810	22536974
SAG0811	22536975
SAG0812	22536976
SAG0814	22536978
SAG0815	22536979
SAG0816	22536980
SAG0817	22536981

Locus Tag	Gi
SAG0818	22536982
SAG0819	22536983
SAG0820	22536984
SAG0821	22536985
SAG0822	22536986
SAG0823	22536987
SAG0824	22536988
SAG0825	22536989
SAG0826	22536990
SAG0827	22536991
SAG0828	22536992
SAG0830	22536994
SAG0831	22536995
SAG0833	22536997
SAG0834	22536998
SAG0835	22536999
SAG0836	22537000
SAG0837	22537001
SAG0838	22537002
SAG0839	22537003
SAG0840	22537004
SAG0841	22537005
SAG0842	22537006
SAG0843	22537007
SAG0844	22537008
SAG0845	22537009
SAG0846	22537010
SAG0847	22537011
SAG0848	22537012
SAG0849	22537013
SAG0850	22537014
SAG0851	22537015
SAG0852	22537016
SAG0853	22537017
SAG0854	22537018
SAG0856	22537019
SAG0857	22537020
SAG0858	22537021
SAG0859	22537022
SAG0860	22537023
SAG0861	22537024
SAG0862	22537025
SAG0863	22537026
SAG0864	22537027
SAG0866	22537029
SAG0867	22537030
SAG0868	22537031
SAG0869	22537032
SAG0870	22537033
SAG0871	22537034
SAG0873	22537036
SAG0874	22537037
SAG0875	22537038
SAG0876	22537039
SAG0877	22537040
SAG0878	22537041
SAG0879	22537042
SAG0880	22537043
SAG0881	22537044
SAG0882	22537045
SAG0883	22537046
SAG0884	22537047
SAG0885	22537048
SAG0886	22537049
SAG0887	22537050
SAG0888	22537051

Locus Tag	Gi
SAG0889	22537052
SAG0890	22537053
SAG0891	22537054
SAG0892	22537055
SAG0893	22537056
SAG0894	22537057
SAG0895	22537058
SAG0896	22537059
SAG0897	22537060
SAG0905	22537068
SAG0906	22537069
SAG0907	22537070
SAG0908	22537071
SAG0909	22537072
SAG0910	22537073
SAG0911	22537074
SAG0912	22537075
SAG0913	22537076
SAG0914	22537077
SAG0938	22537099
SAG0939	22537100
SAG0940	22537101
SAG0941	22537102
SAG0942	22537103
SAG0944	22537105
SAG0946	22537107
SAG0947	22537108
SAG0948	22537109
SAG0949	22537110
SAG0950	161485614
SAG0951	22537112
SAG0952	22537113
SAG0953	22537114
SAG0954	22537115
SAG0955	22537116
SAG0956	22537117
SAG0957	22537118
SAG0958	22537119
SAG0959	22537120
SAG0960	22537121
SAG0961	22537122
SAG0962	22537123
SAG0963	22537124
SAG0964	22537125
SAG0967	22537128
SAG0968	22537129
SAG0969	22537130
SAG0970	22537131
SAG0971	22537132
SAG0974	22537134
SAG0975	22537135
SAG0976	22537136
SAG0977	22537137
SAG0978	22537138
SAG0979	22537139
SAG0980	22537140
SAG0981	22537141
SAG0982	22537142
SAG0983	22537143
SAG0984	22537144
SAG0985	22537145
SAG0986	22537146
SAG0987	22537147
SAG0988	22537148
SAG0989	22537149
SAG0990	22537150

Locus Tag	GI
SAG0991	22537151
SAG0992	22537152
SAG0993	22537153
SAG0994	22537154
SAG0995	22537155
SAG0996	22537156
SAG0997	22537157
SAG0998	22537158
SAG0999	22537159
SAG1000	22537160
SAG1001	22537161
SAG1002	22537162
SAG1003	22537163
SAG1004	22537164
SAG1005	22537165
SAG1006	22537166
SAG1007	22537167
SAG1008	22537168
SAG1009	22537169
SAG1010	22537170
SAG1011	22537171
SAG1012	22537172
SAG1013	22537173
SAG1014	22537174
SAG1015	22537175
SAG1016	22537176
SAG1017	22537177
SAG1027	22537186
SAG1032	22537191
SAG1033	22537192
SAG1034	22537193
SAG1035	22537194
SAG1036	22537195
SAG1037	22537196
SAG1038	22537197
SAG1039	22537198
SAG1040	22537199
SAG1041	22537200
SAG1042	22537201
SAG1043	22537202
SAG1044	22537203
SAG1045	22537204
SAG1046	22537205
SAG1047	22537206
SAG1048	22537207
SAG1049	22537208
SAG1051	22537209
SAG1052	22537210
SAG1054	22537212
SAG1055	22537213
SAG1056	22537214
SAG1057	22537215
SAG1058	22537216
SAG1059	22537217
SAG1060	22537218
SAG1061	22537219
SAG1062	22537220
SAG1063	22537221
SAG1064	22537222
SAG1065	22537223
SAG1066	22537224
SAG1070	22537228
SAG1071	22537229
SAG1072	22537230
SAG1073	22537231
SAG1074	22537232

Locus Tag	GI
SAG1075	22537233
SAG1076	22537234
SAG1077	22537235
SAG1078	22537236
SAG1079	22537237
SAG1081	22537239
SAG1082	22537240
SAG1083	22537241
SAG1084	22537242
SAG1085	22537243
SAG1086	22537244
SAG1087	22537245
SAG1088	22537246
SAG1089	22537247
SAG1090	22537248
SAG1091	22537249
SAG1092	22537250
SAG1093	22537251
SAG1094	22537252
SAG1095	22537253
SAG1096	22537254
SAG1097	22537255
SAG1098	22537256
SAG1099	22537257
SAG1100	22537258
SAG1101	22537259
SAG1102	22537260
SAG1103	22537261
SAG1104	22537262
SAG1105	22537263
SAG1106	22537264
SAG1107	22537265
SAG1108	22537266
SAG1109	22537267
SAG1110	22537268
SAG1111	22537269
SAG1112	22537270
SAG1113	22537271
SAG1114	22537272
SAG1115	22537273
SAG1116	22537274
SAG1117	22537275
SAG1118	22537276
SAG1119	22537277
SAG1120	22537278
SAG1124	22537282
SAG1125	22537283
SAG1126	22537284
SAG1130	22537288
SAG1131	22537289
SAG1132	22537290
SAG1134	22537292
SAG1135	22537293
SAG1136	22537294
SAG1137	22537295
SAG1138	22537296
SAG1139	22537297
SAG1140	22537298
SAG1141	22537299
SAG1142	22537300
SAG1143	22537301
SAG1144	22537302
SAG1145	22537303
SAG1146	22537304
SAG1147	22537305
SAG1148	22537306

Locus Tag	GI
SAG1149	22537307
SAG1150	22537308
SAG1151	22537309
SAG1152	22537310
SAG1153	22537311
SAG1154	22537312
SAG1155	22537313
SAG1156	22537314
SAG1157	22537315
SAG1158	22537316
SAG1159	22537317
SAG1160	22537318
SAG1161	22537319
SAG1162	22537320
SAG1163	22537321
SAG1170	22537328
SAG1171	22537329
SAG1172	22537330
SAG1173	22537331
SAG1174	22537332
SAG1175	22537333
SAG1176	22537334
SAG1178	22537336
SAG1179	22537337
SAG1180	22537338
SAG1181	22537339
SAG1182	22537340
SAG1183	22537341
SAG1184	22537342
SAG1185	22537343
SAG1186	22537344
SAG1187	22537345
SAG1188	22537346
SAG1189	22537347
SAG1190	22537348
SAG1191	22537349
SAG1192	22537350
SAG1193	22537351
SAG1194	22537352
SAG1195	22537353
SAG1196	22537354
SAG1197	22537355
SAG1198	22537356
SAG1199	22537357
SAG1200	22537358
SAG1201	22537359
SAG1202	22537360
SAG1203	22537361
SAG1204	22537362
SAG1205	22537363
SAG1206	22537364
SAG1208	22537366
SAG1209	22537367
SAG1210	22537368
SAG1211	22537369
SAG1212	22537370
SAG1213	22537371
SAG1214	22537372
SAG1215	22537373
SAG1216	22537374
SAG1218	22537375
SAG1219	22537376
SAG1220	22537377
SAG1222	22537378
SAG1223	22537379
SAG1224	22537380

Locus Tag	GI
SAG1225	22537381
SAG1226	22537382
SAG1227	22537383
SAG1228	22537384
SAG1229	22537385
SAG1230	22537386
SAG1233	22537387
SAG1234	22537388
SAG1237	22537390
SAG1241	22537393
SAG1243	22537394
SAG1244	22537395
SAG1245	22537396
SAG1246	22537397
SAG1247	22537398
SAG1301	22537448
SAG1302	22537449
SAG1303	22537450
SAG1305	22537452
SAG1306	22537453
SAG1307	22537454
SAG1310	22537457
SAG1311	22537458
SAG1312	22537459
SAG1314	22537461
SAG1315	22537462
SAG1316	22537463
SAG1317	22537464
SAG1318	22537465
SAG1319	22537466
SAG1320	22537467
SAG1321	22537468
SAG1322	22537469
SAG1323	22537470
SAG1324	22537471
SAG1325	22537472
SAG1326	22537473
SAG1327	22537474
SAG1328	22537475
SAG1329	22537476
SAG1332	22537479
SAG1333	22537480
SAG1334	22537481
SAG1335	22537482
SAG1336	22537483
SAG1337	22537484
SAG1338	22537485
SAG1339	22537486
SAG1340	22537487
SAG1341	22537488
SAG1342	22537489
SAG1343	22537490
SAG1344	22537491
SAG1345	22537492
SAG1346	22537493
SAG1347	22537494
SAG1348	22537495
SAG1349	22537496
SAG1350	22537497
SAG1351	22537498
SAG1352	22537499
SAG1353	22537500
SAG1354	22537501
SAG1355	22537502
SAG1356	22537503
SAG1357	22537504

Locus Tag	Gi
SAG1358	22537505
SAG1359	22537506
SAG1360	22537507
SAG1361	22537508
SAG1362	22537509
SAG1363	22537510
SAG1364	22537511
SAG1365	22537512
SAG1366	22537513
SAG1367	22537514
SAG1368	22537515
SAG1369	22537516
SAG1370	22537517
SAG1371	22537518
SAG1372	22537519
SAG1373	22537520
SAG1374	22537521
SAG1375	22537522
SAG1376	22537523
SAG1377	22537524
SAG1378	22537525
SAG1379	22537526
SAG1380	22537527
SAG1381	22537528
SAG1382	22537529
SAG1383	22537530
SAG1384	22537531
SAG1385	22537532
SAG1386	22537533
SAG1387	22537534
SAG1388	22537535
SAG1389	22537536
SAG1390	22537537
SAG1391	22537538
SAG1392	22537539
SAG1393	22537540
SAG1394	22537541
SAG1395	22537542
SAG1396	22537543
SAG1397	22537544
SAG1398	22537545
SAG1399	22537546
SAG1400	22537547
SAG1401	22537548
SAG1402	22537549
SAG1403	22537550
SAG1410	22537556
SAG1411	22537557
SAG1412	22537558
SAG1413	22537559
SAG1414	22537560
SAG1415	22537561
SAG1416	22537562
SAG1417	22537563
SAG1418	22537564
SAG1419	22537565
SAG1420	22537566
SAG1421	22537567
SAG1422	22537568
SAG1423	22537569
SAG1424	22537570
SAG1425	22537571
SAG1426	22537572
SAG1427	22537573
SAG1428	22537574
SAG1429	22537575

Locus Tag	Gi
SAG1430	22537576
SAG1431	22537577
SAG1432	22537578
SAG1433	22537579
SAG1434	22537580
SAG1435	22537581
SAG1436	22537582
SAG1438	22537584
SAG1439	22537585
SAG1440	22537586
SAG1441	22537587
SAG1442	22537588
SAG1443	22537589
SAG1444	22537590
SAG1447	22537592
SAG1448	22537593
SAG1449	22537594
SAG1450	22537595
SAG1451	22537596
SAG1452	22537597
SAG1453	22537598
SAG1454	22537599
SAG1455	22537600
SAG1459	22537603
SAG1460	22537604
SAG1461	22537605
SAG1462	22537606
SAG1464	22537607
SAG1465	22537608
SAG1466	22537609
SAG1467	22537610
SAG1469	22537612
SAG1470	161485613
SAG1472	22537615
SAG1473	22537616
SAG1474	22537617
SAG1475	22537618
SAG1476	22537619
SAG1477	22537620
SAG1478	22537621
SAG1479	22537622
SAG1480	22537623
SAG1481	22537624
SAG1482	22537625
SAG1483	22537626
SAG1485	22537628
SAG1486	22537629
SAG1487	22537630
SAG1488	22537631
SAG1489	22537632
SAG1490	22537633
SAG1491	22537634
SAG1495	22537638
SAG1498	22537641
SAG1499	22537642
SAG1500	22537643
SAG1501	22537644
SAG1502	22537645
SAG1505	22537648
SAG1506	22537649
SAG1507	22537650
SAG1508	22537651
SAG1509	22537652
SAG1510	22537653
SAG1511	22537654
SAG1512	22537655

Locus Tag	Gi
SAG1513	22537656
SAG1514	22537657
SAG1515	22537658
SAG1516	22537659
SAG1517	22537660
SAG1518	22537661
SAG1519	22537662
SAG1520	22537663
SAG1521	22537664
SAG1522	22537665
SAG1523	22537666
SAG1524	22537667
SAG1528	22537671
SAG1529	22537672
SAG1530	22537673
SAG1531	22537674
SAG1532	22537675
SAG1533	22537676
SAG1534	22537677
SAG1535	22537678
SAG1536	22537679
SAG1537	22537680
SAG1538	22537681
SAG1540	22537683
SAG1541	22537684
SAG1542	22537685
SAG1544	22537686
SAG1545	22537687
SAG1546	22537688
SAG1547	22537689
SAG1551	22537693
SAG1552	22537694
SAG1553	22537695
SAG1554	22537696
SAG1555	22537697
SAG1556	22537698
SAG1557	22537699
SAG1558	22537700
SAG1559	22537701
SAG1561	22537703
SAG1562	22537704
SAG1563	22537705
SAG1564	22537706
SAG1565	22537707
SAG1566	22537708
SAG1567	22537709
SAG1569	22537710
SAG1572	22537713
SAG1573	22537714
SAG1574	22537715
SAG1575	22537716
SAG1577	22537717
SAG1578	22537718
SAG1579	22537719
SAG1580	22537720
SAG1581	22537721
SAG1583	22537723
SAG1585	22537725
SAG1586	22537726
SAG1587	22537727
SAG1588	22537728
SAG1589	22537729
SAG1590	22537730
SAG1591	22537731
SAG1592	22537732
SAG1593	22537733

Locus Tag	Gi
SAG1594	22537734
SAG1595	22537735
SAG1596	22537736
SAG1597	22537737
SAG1598	22537738
SAG1599	22537739
SAG1600	22537740
SAG1601	22537741
SAG1602	22537742
SAG1603	22537743
SAG1604	22537744
SAG1605	22537745
SAG1606	22537746
SAG1607	22537747
SAG1608	22537748
SAG1609	22537749
SAG1610	22537750
SAG1611	22537751
SAG1612	22537752
SAG1613	22537753
SAG1614	22537754
SAG1615	22537755
SAG1616	22537756
SAG1618	22537758
SAG1620	22537760
SAG1621	22537761
SAG1622	22537762
SAG1623	22537763
SAG1624	22537764
SAG1625	22537765
SAG1626	22537766
SAG1627	22537767
SAG1628	22537768
SAG1629	22537769
SAG1630	22537770
SAG1631	22537771
SAG1632	22537772
SAG1633	22537773
SAG1634	22537774
SAG1635	22537775
SAG1637	22537777
SAG1638	22537778
SAG1639	22537779
SAG1640	22537780
SAG1641	22537781
SAG1642	22537782
SAG1643	22537783
SAG1645	22537785
SAG1647	22537787
SAG1648	22537788
SAG1650	22537790
SAG1651	22537791
SAG1652	22537792
SAG1653	22537793
SAG1654	22537794
SAG1655	22537795
SAG1656	22537796
SAG1657	22537797
SAG1658	22537798
SAG1659	22537799
SAG1660	22537800
SAG1661	22537801
SAG1662	22537802
SAG1663	22537803
SAG1664	22537804
SAG1665	22537805

Locus Tag	Gi
SAG1666	22537806
SAG1667	22537807
SAG1668	22537808
SAG1669	22537809
SAG1670	22537810
SAG1671	22537811
SAG1672	22537812
SAG1673	22537813
SAG1674	22537814
SAG1675	22537815
SAG1676	22537816
SAG1677	22537817
SAG1678	22537818
SAG1679	22537819
SAG1680	22537820
SAG1681	22537821
SAG1682	22537822
SAG1683	22537823
SAG1684	22537824
SAG1685	22537825
SAG1686	22537826
SAG1687	22537827
SAG1688	22537828
SAG1689	22537829
SAG1690	22537830
SAG1691	22537831
SAG1692	22537832
SAG1693	22537833
SAG1694	22537834
SAG1695	22537835
SAG1704	22537844
SAG1706	22537845
SAG1707	22537846
SAG1709	22537848
SAG1710	22537849
SAG1711	22537850
SAG1712	22537851
SAG1713	22537852
SAG1714	22537853
SAG1715	22537854
SAG1716	22537855
SAG1717	22537856
SAG1719	22537858
SAG1720	22537859
SAG1721	22537860
SAG1722	22537861
SAG1723	22537862
SAG1724	22537863
SAG1725	22537864
SAG1726	22537865
SAG1727	22537866
SAG1728	22537867
SAG1729	22537868
SAG1730	22537869
SAG1731	22537870
SAG1732	22537871
SAG1733	22537872
SAG1734	22537873
SAG1735	22537874
SAG1736	22537875
SAG1737	22537876
SAG1738	22537877
SAG1739	22537878
SAG1740	22537879
SAG1741	22537880
SAG1742	22537881

Locus Tag	Gi
SAG1743	22537882
SAG1744	22537883
SAG1745	22537884
SAG1747	22537886
SAG1748	22537887
SAG1749	22537888
SAG1750	22537889
SAG1751	22537890
SAG1752	22537891
SAG1753	22537892
SAG1754	22537893
SAG1756	22537895
SAG1757	22537896
SAG1758	22537897
SAG1759	22537898
SAG1760	22537899
SAG1761	22537900
SAG1762	22537901
SAG1763	22537902
SAG1764	22537903
SAG1765	22537904
SAG1766	22537905
SAG1767	22537906
SAG1768	22537907
SAG1769	22537908
SAG1770	22537909
SAG1771	22537910
SAG1772	22537911
SAG1773	22537912
SAG1774	22537913
SAG1775	22537914
SAG1776	22537915
SAG1777	22537916
SAG1778	22537917
SAG1779	22537918
SAG1781	22537920
SAG1782	22537921
SAG1783	22537922
SAG1784	22537923
SAG1785	22537924
SAG1786	22537925
SAG1787	22537926
SAG1788	22537927
SAG1789	22537928
SAG1790	22537929
SAG1791	22537930
SAG1792	22537931
SAG1793	22537932
SAG1794	22537933
SAG1795	22537934
SAG1796	22537935
SAG1797	22537936
SAG1799	22537938
SAG1800	22537939
SAG1801	22537940
SAG1802	22537941
SAG1803	22537942
SAG1804	22537943
SAG1805	22537944
SAG1806	22537945
SAG1807	22537946
SAG1808	22537947
SAG1809	22537948
SAG1810	22537949
SAG1811	22537950
SAG1812	22537951

Locus Tag	Gi
SAG1813	22537952
SAG1814	22537953
SAG1815	22537954
SAG1816	22537955
SAG1818	22537957
SAG1819	22537958
SAG1820	22537959
SAG1821	22537960
SAG1822	22537961
SAG1823	22537962
SAG1824	22537963
SAG1825	22537964
SAG1826	22537965
SAG1827	22537966
SAG1828	22537967
SAG1829	22537968
SAG1830	22537969
SAG1831	22537970
SAG1832	22537971
SAG1833	22537972
SAG1834	22537973
SAG1863	22538002
SAG1873	22538011
SAG1887	22538025
SAG1888	22538026
SAG1889	22538027
SAG1890	22538028
SAG1891	22538029
SAG1894	22538032
SAG1895	22538033
SAG1896	22538034
SAG1909	22538047
SAG1910	22538048
SAG1911	22538049
SAG1912	22538050
SAG1913	22538051
SAG1914	22538052
SAG1915	22538053
SAG1916	22538054
SAG1917	22538055
SAG1918	22538056
SAG1919	22538057
SAG1920	22538058
SAG1921	22538059
SAG1922	22538060
SAG1923	22538061
SAG1924	22538062
SAG1925	22538063
SAG1926	22538064
SAG1927	22538065
SAG1928	22538066
SAG1929	22538067
SAG1930	22538068
SAG1931	22538069
SAG1932	22538070
SAG1933	22538071
SAG1934	22538072
SAG1935	22538073
SAG1936	22538074
SAG1938	22538075
SAG1939	22538076
SAG1940	22538077
SAG1941	22538078
SAG1942	22538079
SAG1943	22538080
SAG1944	22538081

Locus Tag	Gi
SAG1945	22538082
SAG1946	22538083
SAG1947	22538084
SAG1948	22538085
SAG1949	22538086
SAG1950	22538087
SAG1951	22538088
SAG1952	22538089
SAG1954	22538091
SAG1958	22538094
SAG1959	22538095
SAG1960	22538096
SAG1961	22538097
SAG1962	22538098
SAG1963	22538099
SAG1964	22538100
SAG1965	22538101
SAG1966	22538102
SAG1967	22538103
SAG1968	22538104
SAG1969	22538105
SAG1970	22538106
SAG1971	22538107
SAG1972	22538108
SAG1973	22538109
SAG1974	22538110
SAG1975	22538111
SAG1976	22538112
SAG1977	22538113
SAG1978	22538114
SAG1979	22538115
SAG1980	22538116
SAG1982	22538118
SAG1983	22538119
SAG1984	22538120
SAG1985	22538121
SAG2026	22538161
SAG2027	22538162
SAG2029	22538164
SAG2030	22538165
SAG2031	22538166
SAG2032	22538167
SAG2033	22538168
SAG2034	22538169
SAG2035	22538170
SAG2037	22538172
SAG2038	22538173
SAG2039	22538174
SAG2040	22538175
SAG2041	22538176
SAG2042	22538177
SAG2043	22538178
SAG2045	22538180
SAG2046	22538181
SAG2047	22538182
SAG2048	22538183
SAG2049	22538184
SAG2050	22538185
SAG2051	22538186
SAG2053	22538188
SAG2054	22538189
SAG2055	22538190
SAG2056	22538191
SAG2057	22538192
SAG2058	22538193
SAG2059	22538194

Locus Tag	Gi
SAG2062	22538197
SAG2064	22538199
SAG2066	22538201
SAG2067	22538202
SAG2068	22538203
SAG2069	22538204
SAG2070	22538205
SAG2071	22538206
SAG2072	22538207
SAG2073	22538208
SAG2074	22538209
SAG2075	22538210
SAG2076	22538211
SAG2077	22538212
SAG2078	22538213
SAG2079	22538214
SAG2080	22538215
SAG2081	22538216
SAG2082	22538217
SAG2083	22538218
SAG2084	22538219
SAG2085	22538220
SAG2086	22538221
SAG2087	22538222
SAG2089	22538224
SAG2090	22538225
SAG2091	22538226
SAG2092	22538227
SAG2093	22538228
SAG2095	22538229
SAG2096	22538230
SAG2097	22538231
SAG2098	22538232
SAG2100	22538234
SAG2101	22538235
SAG2102	22538236
SAG2103	22538237
SAG2104	22538238
SAG2105	22538239
SAG2106	22538240
SAG2107	22538241
SAG2108	22538242
SAG2109	22538243
SAG2110	22538244
SAG2111	22538245
SAG2121	22538255
SAG2122	22538256
SAG2123	22538257
SAG2124	22538258
SAG2125	22538259
SAG2126	22538260
SAG2127	22538261
SAG2128	22538262
SAG2129	22538263
SAG2130	22538264
SAG2131	22538265
SAG2132	22538266
SAG2133	22538267
SAG2134	22538268
SAG2135	22538269
SAG2136	22538270
SAG2137	22538271
SAG2138	22538272
SAG2139	22538273
SAG2140	22538274
SAG2141	22538275

Locus Tag	Gi
SAG2142	22538276
SAG2143	22538277
SAG2144	22538278
SAG2145	22538279
SAG2146	22538280
SAG2147	22538281
SAG2148	22538282
SAG2149	22538283
SAG2150	22538284
SAG2151	22538285
SAG2152	22538286
SAG2153	22538287
SAG2154	22538288
SAG2155	22538289
SAG2156	22538290
SAG2157	22538291
SAG2158	22538292
SAG2159	22538293
SAG2160	22538294
SAG2161	22538295
SAG2162	22538296
SAG2163	22538297
SAG2164	22538298
SAG2165	22538299
SAG2166	22538300
SAG2167	22538301
SAG2168	22538302
SAG2169	22538303
SAG2170	22538304
SAG2171	22538305
SAG2172	22538306
SAG2173	22538307
SAG2174	22538308
SAG2175	22538309

9.10.2 Eight Genome Dataset Core Genome

Locus Tag	Gi
SAG0001	22536186
SAG0002	22536187
SAG0006	22536191
SAG0007	22536192
SAG0011	22536196
SAG0012	22536197
SAG0013	22536198
SAG0014	22536199
SAG0015	22536200
SAG0016	22536201
SAG0017	22536202
SAG0018	22536203
SAG0019	22536204
SAG0021	22536206
SAG0022	22536207
SAG0024	22536209
SAG0025	22536210
SAG0027	22536212
SAG0028	22536213
SAG0030	22536215
SAG0031	22536216
SAG0032	22536217
SAG0033	22536218
SAG0034	22536219
SAG0035	22536220
SAG0036	22536221
SAG0037	22536222
SAG0038	22536223
SAG0039	22536224
SAG0040	22536225
SAG0041	22536226
SAG0042	22536227
SAG0043	22536228
SAG0044	22536229
SAG0045	22536230
SAG0047	22536232
SAG0049	22536234
SAG0050	22536235
SAG0052	22536237
SAG0054	22536239
SAG0055	22536240
SAG0056	22536241
SAG0058	22536243
SAG0060	22536245
SAG0061	22536246
SAG0062	22536247
SAG0063	22536248
SAG0064	22536249
SAG0065	22536250
SAG0066	22536251
SAG0067	22536252
SAG0068	22536253
SAG0070	22536255
SAG0072	22536257
SAG0073	22536258
SAG0074	22536259
SAG0075	22536260
SAG0076	22536261
SAG0077	22536262
SAG0078	22536263
SAG0079	22536264
SAG0085	22536270
SAG0086	22536271
SAG0089	22536274
SAG0090	22536275
SAG0091	22536276
SAG0092	22536277
SAG0094	22536279
SAG0095	22536280
SAG0096	22536281
SAG0097	22536282
SAG0098	22536283
SAG0099	22536284
SAG0100	22536285
SAG0111	22536296
SAG0112	22536297
SAG0113	22536298
SAG0115	22536300
SAG0116	22536301
SAG0117	22536302
SAG0119	22536304
SAG0121	22536306
SAG0132	22536317
SAG0136	22536321
SAG0138	22536323
SAG0139	22536324
SAG0140	22536325
SAG0141	22536326
SAG0142	22536327
SAG0143	22536328
SAG0144	22536329
SAG0145	22536330
SAG0148	22536333
SAG0150	22536335
SAG0151	22536336
SAG0152	22536337
SAG0154	22536339
SAG0155	22536340
SAG0158	22536342
SAG0159	22536343
SAG0160	22536344
SAG0161	22536345
SAG0162	22536346
SAG0164	22536348
SAG0165	22536349
SAG0167	22536351
SAG0168	22536352
SAG0169	22536353
SAG0172	22536356
SAG0173	22536357
SAG0174	22536358
SAG0176	22536360
SAG0177	22536361
SAG0178	22536362
SAG0179	22536363
SAG0180	22536364
SAG0181	22536365
SAG0182	22536366
SAG0183	22536367
SAG0184	22536368
SAG0185	22536369
SAG0187	22536371
SAG0188	22536372
SAG0189	22536373
SAG0190	22536374
SAG0191	22536375
SAG0194	22536378
SAG0196	22536380
SAG0197	22536381
SAG0198	22536382
SAG0199	22536383
SAG0200	22536384
SAG0201	22536385
SAG0202	22536386
SAG0203	22536387
SAG0204	22536388
SAG0205	22536389
SAG0206	22536390
SAG0207	22536391
SAG0208	22536392
SAG0209	22536393
SAG0210	22536394
SAG0211	22536395
SAG0214	22536398
SAG0215	22536399
SAG0220	22536404
SAG0239	22536423
SAG0241	22536425
SAG0242	22536426
SAG0244	22536428
SAG0252	22536436
SAG0254	22536438
SAG0255	22536439
SAG0256	22536440
SAG0257	22536441
SAG0258	22536442
SAG0264	22536448
SAG0265	22536449
SAG0266	22536450
SAG0268	22536452
SAG0269	22536453
SAG0270	22536454
SAG0271	22536455
SAG0272	22536456
SAG0273	22536457
SAG0274	22536458
SAG0275	22536459
SAG0276	22536460
SAG0277	22536461
SAG0279	22536463
SAG0280	22536464
SAG0281	22536465
SAG0283	22536467
SAG0284	22536468
SAG0285	161485617
SAG0286	22536470
SAG0287	22536471
SAG0288	22536472
SAG0289	22536473
SAG0290	22536474
SAG0291	22536475
SAG0292	22536476
SAG0293	22536477
SAG0294	22536478
SAG0295	22536479
SAG0296	22536480
SAG0297	22536481
SAG0298	22536482
SAG0299	22536483
SAG0300	22536484
SAG0304	22536488
SAG0305	22536489
SAG0306	22536490
SAG0312	22536495
SAG0313	22536496
SAG0314	22536497
SAG0315	22536498
SAG0316	22536499
SAG0317	22536500
SAG0318	22536501
SAG0319	22536502
SAG0320	22536503
SAG0321	22536504
SAG0322	22536505
SAG0323	22536506
SAG0324	22536507
SAG0325	22536508
SAG0326	22536509
SAG0327	22536510
SAG0328	22536511
SAG0329	22536512
SAG0330	22536513
SAG0331	22536514
SAG0332	22536515
SAG0334	22536517
SAG0335	22536518
SAG0336	22536519
SAG0338	22536521
SAG0339	22536522
SAG0340	22536523
SAG0342	22536525
SAG0343	22536526
SAG0344	22536527
SAG0345	22536528
SAG0346	22536529
SAG0347	22536530
SAG0348	22536531
SAG0349	22536532
SAG0350	22536533
SAG0351	22536534
SAG0352	22536535
SAG0353	22536536
SAG0354	22536537
SAG0355	22536538
SAG0356	22536539
SAG0357	22536540
SAG0358	22536541
SAG0359	22536542
SAG0360	22536543
SAG0363	22536546
SAG0364	22536547
SAG0365	22536548
SAG0366	22536549
SAG0367	22536550
SAG0368	22536551
SAG0369	22536552
SAG0370	22536553
SAG0371	22536554
SAG0373	22536556
SAG0374	22536557
SAG0375	22536558
SAG0376	22536559
SAG0377	22536560
SAG0378	161485616
SAG0379	22536562
SAG0380	22536563
SAG0382	22536565
SAG0383	22536566
SAG0385	22536568
SAG0386	22536569
SAG0387	22536570
SAG0388	22536571
SAG0389	22536572
SAG0390	22536573

Locus Tag	GI
SAG0391	22536574
SAG0392	22536575
SAG0393	22536576
SAG0394	22536577
SAG0395	22536578
SAG0397	22536580
SAG0398	22536581
SAG0400	22536583
SAG0402	22536585
SAG0403	22536586
SAG0408	22536591
SAG0409	22536592
SAG0410	22536593
SAG0411	22536594
SAG0412	22536595
SAG0413	22536596
SAG0417	22536600
SAG0418	22536601
SAG0419	22536602
SAG0420	22536603
SAG0421	22536604
SAG0422	22536605
SAG0423	22536606
SAG0431	22536614
SAG0432	22536615
SAG0443	22536623
SAG0445	22536625
SAG0451	22536631
SAG0452	22536632
SAG0454	22536634
SAG0455	22536635
SAG0457	22536636
SAG0458	22536637
SAG0459	22536638
SAG0460	22536639
SAG0461	22536640
SAG0462	22536641
SAG0464	22536643
SAG0465	22536644
SAG0467	22536646
SAG0468	22536647
SAG0469	22536648
SAG0470	22536649
SAG0471	22536650
SAG0472	22536651
SAG0473	22536652
SAG0474	22536653
SAG0475	22536654
SAG0476	22536655
SAG0477	22536656
SAG0478	22536657
SAG0479	22536658
SAG0480	22536659
SAG0481	22536660
SAG0484	22536663
SAG0486	22536665
SAG0487	22536666
SAG0488	22536667
SAG0490	22536669
SAG0493	22536672
SAG0494	22536673
SAG0495	22536674
SAG0496	22536675
SAG0497	22536676
SAG0498	22536677
SAG0499	22536678

Locus Tag	GI
SAG0500	22536679
SAG0501	22536680
SAG0503	22536682
SAG0506	22536685
SAG0508	22536687
SAG0509	22536688
SAG0510	22536689
SAG0511	22536690
SAG0512	22536691
SAG0513	22536692
SAG0516	22536695
SAG0517	22536696
SAG0519	22536697
SAG0520	22536698
SAG0521	22536699
SAG0522	22536700
SAG0524	22536702
SAG0525	22536703
SAG0526	22536704
SAG0527	22536705
SAG0528	22536706
SAG0530	22536708
SAG0531	22536709
SAG0532	22536710
SAG0533	22536711
SAG0534	22536712
SAG0536	22536714
SAG0537	22536715
SAG0538	22536716
SAG0544	22536722
SAG0606	22536782
SAG0610	22536785
SAG0612	22536786
SAG0613	22536787
SAG0614	22536788
SAG0616	22536790
SAG0619	22536792
SAG0621	22536794
SAG0622	22536795
SAG0623	22536796
SAG0624	22536797
SAG0625	22536798
SAG0626	22536799
SAG0627	22536800
SAG0628	22536801
SAG0630	22536803
SAG0631	22536804
SAG0632	22536805
SAG0633	22536806
SAG0657	22536824
SAG0658	22536825
SAG0660	22536827
SAG0663	22536830
SAG0665	22536832
SAG0666	22536833
SAG0667	22536834
SAG0668	22536835
SAG0669	22536836
SAG0670	22536837
SAG0671	22536838
SAG0672	22536839
SAG0673	22536840
SAG0684	22536849
SAG0686	22536851
SAG0690	22536855
SAG0691	22536856

Locus Tag	GI
SAG0692	22536857
SAG0695	22536859
SAG0697	22536861
SAG0698	22536862
SAG0699	22536863
SAG0700	22536864
SAG0701	22536865
SAG0702	22536866
SAG0703	22536867
SAG0704	22536868
SAG0705	22536869
SAG0706	22536870
SAG0707	22536871
SAG0708	22536872
SAG0709	22536873
SAG0711	22536875
SAG0712	22536876
SAG0714	22536878
SAG0715	22536879
SAG0716	22536880
SAG0718	22536882
SAG0719	22536883
SAG0720	22536884
SAG0721	22536885
SAG0722	22536886
SAG0723	22536887
SAG0724	22536888
SAG0725	22536889
SAG0726	22536890
SAG0727	22536891
SAG0728	22536892
SAG0729	22536893
SAG0730	22536894
SAG0731	22536895
SAG0732	22536896
SAG0733	22536897
SAG0736	22536900
SAG0738	22536902
SAG0739	22536903
SAG0741	22536905
SAG0743	22536907
SAG0744	22536908
SAG0745	22536909
SAG0746	22536910
SAG0747	161485615
SAG0748	22536912
SAG0749	22536913
SAG0750	22536914
SAG0751	22536915
SAG0752	22536916
SAG0753	22536917
SAG0757	22536921
SAG0758	22536922
SAG0759	22536923
SAG0761	22536925
SAG0762	22536926
SAG0763	22536927
SAG0764	22536928
SAG0765	22536929
SAG0766	22536930
SAG0767	22536931
SAG0768	22536932
SAG0770	22536934
SAG0775	22536939
SAG0776	22536940
SAG0777	22536941

Locus Tag	GI
SAG0778	22536942
SAG0779	22536943
SAG0780	22536944
SAG0785	22536949
SAG0786	22536950
SAG0787	22536951
SAG0788	22536952
SAG0789	22536953
SAG0790	22536954
SAG0791	22536955
SAG0793	22536957
SAG0794	22536958
SAG0795	22536959
SAG0796	22536960
SAG0797	22536961
SAG0800	22536964
SAG0809	22536973
SAG0810	22536974
SAG0811	22536975
SAG0812	22536976
SAG0815	22536979
SAG0818	22536982
SAG0819	22536983
SAG0820	22536984
SAG0822	22536986
SAG0823	22536987
SAG0824	22536988
SAG0825	22536989
SAG0827	22536991
SAG0828	22536992
SAG0830	22536994
SAG0831	22536995
SAG0833	22536997
SAG0835	22536999
SAG0837	22537001
SAG0838	22537002
SAG0839	22537003
SAG0840	22537004
SAG0841	22537005
SAG0842	22537006
SAG0843	22537007
SAG0845	22537009
SAG0846	22537010
SAG0848	22537012
SAG0849	22537013
SAG0850	22537014
SAG0851	22537015
SAG0852	22537016
SAG0853	22537017
SAG0856	22537019
SAG0858	22537021
SAG0859	22537022
SAG0860	22537023
SAG0861	22537024
SAG0862	22537025
SAG0863	22537026
SAG0864	22537027
SAG0867	22537030
SAG0868	22537031
SAG0869	22537032
SAG0877	22537040
SAG0879	22537042
SAG0880	22537043
SAG0881	22537044
SAG0882	22537045
SAG0885	22537048

Locus Tag	Gi
SAG0886	22537049
SAG0887	22537050
SAG0888	22537051
SAG0890	22537053
SAG0891	22537054
SAG0892	22537055
SAG0893	22537056
SAG0896	22537059
SAG0897	22537060
SAG0905	22537068
SAG0906	22537069
SAG0910	22537073
SAG0912	22537075
SAG0913	22537076
SAG0940	22537101
SAG0941	22537102
SAG0942	22537103
SAG0944	22537105
SAG0947	22537108
SAG0948	22537109
SAG0949	22537110
SAG0950	161485614
SAG0951	22537112
SAG0952	22537113
SAG0955	22537116
SAG0956	22537117
SAG0957	22537118
SAG0958	22537119
SAG0960	22537121
SAG0961	22537122
SAG0962	22537123
SAG0967	22537128
SAG0968	22537129
SAG0969	22537130
SAG0970	22537131
SAG0971	22537132
SAG0974	22537134
SAG0976	22537136
SAG0978	22537138
SAG0980	22537140
SAG0981	22537141
SAG0982	22537142
SAG0983	22537143
SAG0984	22537144
SAG0985	22537145
SAG0986	22537146
SAG0987	22537147
SAG0988	22537148
SAG0989	22537149
SAG0990	22537150
SAG0991	22537151
SAG0994	22537154
SAG0999	22537159
SAG1004	22537164
SAG1005	22537165
SAG1006	22537166
SAG1008	22537168
SAG1009	22537169
SAG1010	22537170
SAG1012	22537172
SAG1013	22537173
SAG1015	22537175
SAG1016	22537176
SAG1027	22537186
SAG1033	22537192
SAG1034	22537193

Locus Tag	Gi
SAG1035	22537194
SAG1036	22537195
SAG1041	22537200
SAG1042	22537201
SAG1043	22537202
SAG1044	22537203
SAG1045	22537204
SAG1046	22537205
SAG1047	22537206
SAG1048	22537207
SAG1049	22537208
SAG1052	22537210
SAG1054	22537212
SAG1055	22537213
SAG1056	22537214
SAG1058	22537216
SAG1059	22537217
SAG1060	22537218
SAG1061	22537219
SAG1062	22537220
SAG1063	22537221
SAG1064	22537222
SAG1065	22537223
SAG1066	22537224
SAG1070	22537228
SAG1072	22537230
SAG1073	22537231
SAG1074	22537232
SAG1075	22537233
SAG1076	22537234
SAG1077	22537235
SAG1078	22537236
SAG1079	22537237
SAG1081	22537239
SAG1082	22537240
SAG1083	22537241
SAG1084	22537242
SAG1085	22537243
SAG1086	22537244
SAG1087	22537245
SAG1088	22537246
SAG1092	22537250
SAG1093	22537251
SAG1095	22537253
SAG1096	22537254
SAG1097	22537255
SAG1098	22537256
SAG1099	22537257
SAG1100	22537258
SAG1102	22537260
SAG1103	22537261
SAG1104	22537262
SAG1105	22537263
SAG1107	22537265
SAG1108	22537266
SAG1109	22537267
SAG1110	22537268
SAG1111	22537269
SAG1112	22537270
SAG1113	22537271
SAG1114	22537272
SAG1115	22537273
SAG1116	22537274
SAG1117	22537275
SAG1118	22537276
SAG1119	22537277

Locus Tag	Gi
SAG1120	22537278
SAG1124	22537282
SAG1125	22537283
SAG1142	22537300
SAG1143	22537301
SAG1146	22537304
SAG1147	22537305
SAG1148	22537306
SAG1149	22537307
SAG1151	22537309
SAG1153	22537311
SAG1154	22537312
SAG1156	22537314
SAG1157	22537315
SAG1159	22537317
SAG1161	22537319
SAG1162	22537320
SAG1170	22537328
SAG1171	22537329
SAG1172	22537330
SAG1173	22537331
SAG1174	22537332
SAG1175	22537333
SAG1176	22537334
SAG1178	22537336
SAG1179	22537337
SAG1180	22537338
SAG1181	22537339
SAG1182	22537340
SAG1185	22537343
SAG1188	22537346
SAG1190	22537348
SAG1191	22537349
SAG1192	22537350
SAG1193	22537351
SAG1194	22537352
SAG1195	22537353
SAG1196	22537354
SAG1197	22537355
SAG1198	22537356
SAG1199	22537357
SAG1200	22537358
SAG1201	22537359
SAG1202	22537360
SAG1203	22537361
SAG1204	22537362
SAG1205	22537363
SAG1206	22537364
SAG1210	22537368
SAG1211	22537369
SAG1213	22537371
SAG1214	22537372
SAG1215	22537373
SAG1218	22537375
SAG1219	22537376
SAG1220	22537377
SAG1228	22537384
SAG1229	22537385
SAG1230	22537386
SAG1233	22537387
SAG1237	22537390
SAG1243	22537394
SAG1244	22537395
SAG1245	22537396
SAG1302	22537449
SAG1305	22537452

Locus Tag	Gi
SAG1306	22537453
SAG1307	22537454
SAG1311	22537458
SAG1312	22537459
SAG1314	22537461
SAG1316	22537463
SAG1318	22537465
SAG1319	22537466
SAG1320	22537467
SAG1321	22537468
SAG1322	22537469
SAG1323	22537470
SAG1324	22537471
SAG1325	22537472
SAG1326	22537473
SAG1327	22537474
SAG1328	22537475
SAG1334	22537481
SAG1335	22537482
SAG1336	22537483
SAG1337	22537484
SAG1338	22537485
SAG1340	22537487
SAG1341	22537488
SAG1342	22537489
SAG1344	22537491
SAG1345	22537492
SAG1347	22537494
SAG1349	22537496
SAG1350	22537497
SAG1351	22537498
SAG1352	22537499
SAG1354	22537501
SAG1363	22537510
SAG1364	22537511
SAG1365	22537512
SAG1366	22537513
SAG1367	22537514
SAG1368	22537515
SAG1369	22537516
SAG1370	22537517
SAG1371	22537518
SAG1372	22537519
SAG1373	22537520
SAG1376	22537523
SAG1378	22537525
SAG1379	22537526
SAG1380	22537527
SAG1382	22537529
SAG1384	22537531
SAG1390	22537537
SAG1391	22537538
SAG1392	22537539
SAG1394	22537541
SAG1395	22537542
SAG1396	22537543
SAG1397	22537544
SAG1399	22537546
SAG1400	22537547
SAG1401	22537548
SAG1402	22537549
SAG1403	22537550
SAG1410	22537556
SAG1411	22537557
SAG1412	22537558
SAG1413	22537559

Locus Tag	GI
SAG1414	22537560
SAG1415	22537561
SAG1416	22537562
SAG1417	22537563
SAG1418	22537564
SAG1420	22537566
SAG1421	22537567
SAG1426	22537572
SAG1434	22537580
SAG1436	22537582
SAG1448	22537593
SAG1464	22537607
SAG1465	22537608
SAG1466	22537609
SAG1467	22537610
SAG1469	22537612
SAG1470	161485613
SAG1472	22537615
SAG1473	22537616
SAG1474	22537617
SAG1475	22537618
SAG1476	22537619
SAG1477	22537620
SAG1478	22537621
SAG1479	22537622
SAG1481	22537624
SAG1482	22537625
SAG1483	22537626
SAG1485	22537628
SAG1499	22537642
SAG1500	22537643
SAG1507	22537650
SAG1512	22537655
SAG1513	22537656
SAG1515	22537658
SAG1516	22537659
SAG1517	22537660
SAG1518	22537661
SAG1519	22537662
SAG1523	22537666
SAG1524	22537667
SAG1528	22537671
SAG1534	22537677
SAG1535	22537678
SAG1536	22537679
SAG1537	22537680
SAG1538	22537681
SAG1542	22537685
SAG1544	22537686
SAG1545	22537687
SAG1546	22537688
SAG1547	22537689
SAG1551	22537693
SAG1552	22537694
SAG1553	22537695
SAG1554	22537696
SAG1555	22537697
SAG1556	22537698
SAG1557	22537699
SAG1559	22537701
SAG1561	22537703
SAG1562	22537704
SAG1563	22537705
SAG1564	22537706
SAG1567	22537709
SAG1569	22537710

Locus Tag	GI
SAG1572	22537713
SAG1573	22537714
SAG1575	22537716
SAG1577	22537717
SAG1578	22537718
SAG1579	22537719
SAG1580	22537720
SAG1581	22537721
SAG1583	22537723
SAG1587	22537727
SAG1590	22537730
SAG1591	22537731
SAG1592	22537732
SAG1593	22537733
SAG1596	22537736
SAG1597	22537737
SAG1598	22537738
SAG1602	22537742
SAG1603	22537743
SAG1604	22537744
SAG1605	22537745
SAG1606	22537746
SAG1607	22537747
SAG1608	22537748
SAG1609	22537749
SAG1610	22537750
SAG1611	22537751
SAG1612	22537752
SAG1613	22537753
SAG1614	22537754
SAG1615	22537755
SAG1616	22537756
SAG1618	22537758
SAG1620	22537760
SAG1621	22537761
SAG1622	22537762
SAG1623	22537763
SAG1625	22537765
SAG1626	22537766
SAG1627	22537767
SAG1628	22537768
SAG1629	22537769
SAG1638	22537778
SAG1639	22537779
SAG1640	22537780
SAG1643	22537783
SAG1645	22537785
SAG1647	22537787
SAG1648	22537788
SAG1651	22537791
SAG1652	22537792
SAG1653	22537793
SAG1654	22537794
SAG1656	22537796
SAG1657	22537797
SAG1659	22537799
SAG1667	22537807
SAG1668	22537808
SAG1669	22537809
SAG1671	22537811
SAG1672	22537812
SAG1676	22537816
SAG1686	22537826
SAG1687	22537827
SAG1688	22537828
SAG1689	22537829

Locus Tag	GI
SAG1690	22537830
SAG1691	22537831
SAG1693	22537833
SAG1694	22537834
SAG1695	22537835
SAG1706	22537845
SAG1709	22537848
SAG1710	22537849
SAG1711	22537850
SAG1715	22537854
SAG1716	22537855
SAG1717	22537856
SAG1719	22537858
SAG1720	22537859
SAG1721	22537860
SAG1722	22537861
SAG1723	22537862
SAG1724	22537863
SAG1725	22537864
SAG1726	22537865
SAG1727	22537866
SAG1729	22537868
SAG1730	22537869
SAG1731	22537870
SAG1732	22537871
SAG1733	22537872
SAG1735	22537874
SAG1736	22537875
SAG1737	22537876
SAG1738	22537877
SAG1739	22537878
SAG1740	22537879
SAG1741	22537880
SAG1747	22537886
SAG1748	22537887
SAG1750	22537889
SAG1751	22537890
SAG1754	22537893
SAG1756	22537895
SAG1757	22537896
SAG1758	22537897
SAG1759	22537898
SAG1760	22537899
SAG1761	22537900
SAG1762	22537901
SAG1763	22537902
SAG1764	22537903
SAG1766	22537905
SAG1767	22537906
SAG1771	22537910
SAG1772	22537911
SAG1773	22537912
SAG1774	22537913
SAG1775	22537914
SAG1776	22537915
SAG1777	22537916
SAG1778	22537917
SAG1779	22537918
SAG1781	22537920
SAG1782	22537921
SAG1783	22537922
SAG1784	22537923
SAG1786	22537925
SAG1787	22537926
SAG1788	22537927
SAG1789	22537928

Locus Tag	GI
SAG1790	22537929
SAG1791	22537930
SAG1792	22537931
SAG1796	22537935
SAG1797	22537936
SAG1799	22537938
SAG1800	22537939
SAG1801	22537940
SAG1802	22537941
SAG1804	22537943
SAG1806	22537945
SAG1808	22537947
SAG1809	22537948
SAG1810	22537949
SAG1811	22537950
SAG1812	22537951
SAG1813	22537952
SAG1814	22537953
SAG1816	22537955
SAG1819	22537958
SAG1821	22537960
SAG1822	22537961
SAG1823	22537962
SAG1824	22537963
SAG1825	22537964
SAG1826	22537965
SAG1827	22537966
SAG1828	22537967
SAG1829	22537968
SAG1830	22537969
SAG1832	22537971
SAG1833	22537972
SAG1834	22537973
SAG1888	22538026
SAG1889	22538027
SAG1890	22538028
SAG1891	22538029
SAG1894	22538032
SAG1895	22538033
SAG1896	22538034
SAG1909	22538047
SAG1910	22538048
SAG1911	22538049
SAG1913	22538051
SAG1914	22538052
SAG1915	22538053
SAG1916	22538054
SAG1918	22538056
SAG1919	22538057
SAG1920	22538058
SAG1921	22538059
SAG1922	22538060
SAG1923	22538061
SAG1924	22538062
SAG1925	22538063
SAG1926	22538064
SAG1927	22538065
SAG1928	22538066
SAG1929	22538067
SAG1930	22538068
SAG1932	22538070
SAG1934	22538072
SAG1935	22538073
SAG1938	22538075
SAG1939	22538076
SAG1941	22538078

Locus Tag	GI
SAG1942	22538079
SAG1943	22538080
SAG1944	22538081
SAG1945	22538082
SAG1946	22538083
SAG1947	22538084
SAG1948	22538085
SAG1949	22538086
SAG1950	22538087
SAG1951	22538088
SAG1952	22538089
SAG1954	22538091
SAG1958	22538094
SAG1959	22538095
SAG1960	22538096
SAG1961	22538097
SAG1962	22538098
SAG1963	22538099
SAG1964	22538100
SAG1965	22538101
SAG1966	22538102
SAG1968	22538104
SAG1969	22538105
SAG1970	22538106
SAG1972	22538108
SAG1973	22538109
SAG1974	22538110
SAG1975	22538111
SAG1976	22538112
SAG1977	22538113
SAG1978	22538114
SAG1979	22538115
SAG1980	22538116
SAG1982	22538118
SAG1983	22538119
SAG1984	22538120
SAG2032	22538167
SAG2033	22538168
SAG2034	22538169
SAG2035	22538170
SAG2037	22538172
SAG2038	22538173
SAG2039	22538174
SAG2040	22538175
SAG2043	22538178
SAG2046	22538181
SAG2047	22538182
SAG2049	22538184
SAG2050	22538185
SAG2051	22538186
SAG2053	22538188
SAG2055	22538190
SAG2056	22538191
SAG2057	22538192
SAG2058	22538193
SAG2062	22538197
SAG2064	22538199
SAG2066	22538201
SAG2068	22538203
SAG2069	22538204
SAG2070	22538205
SAG2072	22538207
SAG2073	22538208
SAG2075	22538210
SAG2078	22538213
SAG2079	22538214

Locus Tag	GI
SAG2089	22538224
SAG2090	22538225
SAG2092	22538227
SAG2096	22538230
SAG2100	22538234
SAG2101	22538235
SAG2102	22538236
SAG2103	22538237
SAG2104	22538238
SAG2105	22538239
SAG2106	22538240
SAG2107	22538241
SAG2108	22538242
SAG2109	22538243
SAG2110	22538244
SAG2121	22538255
SAG2122	22538256
SAG2123	22538257
SAG2124	22538258
SAG2125	22538259
SAG2126	22538260
SAG2127	22538261
SAG2129	22538263
SAG2131	22538265
SAG2132	22538266
SAG2133	22538267
SAG2134	22538268
SAG2135	22538269
SAG2136	22538270
SAG2137	22538271
SAG2138	22538272
SAG2142	22538276
SAG2143	22538277
SAG2144	22538278
SAG2145	22538279
SAG2146	22538280
SAG2147	22538281
SAG2149	22538283
SAG2150	22538284
SAG2152	22538286
SAG2157	22538291
SAG2158	22538292
SAG2160	22538294
SAG2161	22538295
SAG2162	22538296
SAG2163	22538297
SAG2165	22538299
SAG2166	22538300
SAG2167	22538301
SAG2168	22538302
SAG2170	22538304
SAG2171	22538305
SAG2172	22538306
SAG2173	22538307
SAG2174	22538308
SAG2175	22538309